

Article

Hybrid Machine Learning Models for Predicting Gross CO₂e Balance in Polish Forest Stands: A Tool for Sustainable Forest Carbon Assessment in the Circular Economy

Krzysztof Przybył^{1,*} , Agnieszka A. Pilarska^{2,*}  and Krzysztof Pilarski¹ 

¹ Department of Biosystems Engineering, Poznań University of Life Sciences, Wojska Polskiego 50, 60-627 Poznań, Poland; krzysztof.pilarski@up.poznan.pl

² Department of Hydraulic and Sanitary Engineering, Poznań University of Life Sciences, Piątkowska 94A, 60-649 Poznań, Poland

* Correspondence: krzysztof.przybyl@up.poznan.pl (K.P.); agnieszka.pilarska@up.poznan.pl (A.A.P.); Tel.: +48-61-848-71-56 (K.P.)

Abstract

Forest carbon assessment requires methods that capture the combined effects of stand structure, site conditions, carbon pools, operational emissions, and circular-economy processes. This study aimed to develop and optimize hybrid machine learning models for predicting the gross CO₂e (carbon dioxide equivalent) balance of Polish forest stands using measurable stand- and site-related variables. The research was based on a primary dataset describing forest management in major Polish macroregions in 2020–2024. After data cleaning and preprocessing, multiple machine learning algorithms, including ensemble, boosting, neural, and hybrid models, were trained, validated, and tested. Model performance was assessed using standard regression metrics, overfitting diagnostics, learning curves, and SHAP (Shapley Additive Explanations). Most models achieved high predictive accuracy, with six of ten algorithms reaching R^2 values above 0.90 on the test set. The reduction in strongly correlated variables helped limit multicollinearity and excessive overlap between predictors and the target variable, supporting a more reliable interpretation of model performance. The CatBoost algorithm achieved the highest predictive performance on the test set ($R^2 = 0.948$), while also recording the lowest root mean squared error (RMSE = 152.242). However, the Decision Tree demonstrated the weakest generalization performance ($R^2 = 0.806$) on the test set. SHAP analysis identified tree height as the most influential predictor, followed by tree age, number of trees, species composition, and selected habitat features. The novelty of the study lies in integrating hybrid machine learning, interpretable modelling, and circular-economy-related carbon balance components into a single framework for rapid and operational forest carbon assessment in Polish forest stands.



Academic Editor: Grigorios L. Kyriakopoulos

Received: 13 April 2026

Revised: 1 June 2026

Accepted: 6 June 2026

Published: 22 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: ensemble methods; CatBoost; XGBoost; bioeconomy; greenhouse gas sequestration; forest carbon assessment; Polish forest stands; sustainable forest management; circular economy; carbon inventory; SHAP interpretability

1. Introduction

Forest ecosystems constitute one of the most important components of the terrestrial carbon cycle because they simultaneously accumulate carbon in biomass, soil, and deadwood. They also respond to climatic pressure, biotic disturbances, and management

decisions in ways that vary markedly across space and time [1]. In recent years, however, it has become increasingly evident that the general identification of forests as CO₂ sinks (carbon dioxide) is insufficient for the purposes of contemporary climate policy and forest resource management practice [2]. In reality, the capacity of forest stands to perform a climatic function depends on multiple interacting factors, including species composition, age, site productivity, water conditions, harvesting intensity, drought susceptibility, and the scale of natural and anthropogenic disturbances [2,3]. Research conducted in Central Europe further shows that even regions traditionally regarded as stable carbon sinks may, within a short period, weaken and become temporary sources of emissions under the influence of drought, stand dieback, and increased sanitary felling [1]. Reliable assessment of the forest carbon balance therefore requires a shift from simplified, static descriptions to approaches that integrate ecosystem dynamics, forest management, and environmental variability [2,3]. The development of remote methods and data analytics has also increased the demand for models capable of rapidly and reliably estimating carbon stocks and emissions. Such models should rely on variables measurable in the field and under operational conditions, without repeated and costly direct biomass measurements [4].

The contemporary interpretation of the forest carbon balance must include not only the carbon pool contained in living above-ground and below-ground biomass, but also soil organic carbon stocks, temporal changes in those stocks, carbon flows into wood products, and emissions associated with the harvesting, transport, and processing of raw material [5]. The circular economy (CE) is a system aimed at reducing waste and increasing the value of utilized resources [6]. In forest management, CE's CO₂e balance no longer depends solely on the biological productivity of trees. It requires the integration of technological and logistical processes throughout the entire supply chain to maintain the value of biomass in the economy for as long as possible [7]. Research confirms that the positive impact of forestry on the climate goes beyond the mere absorption of carbon dioxide by ecosystems. It turns out that sustainable resource management is of paramount importance, including the multi-stage use of wood (i.e., its repeated processing across various product life cycles), increasing product durability, and substitution potential, which means directly replacing materials and fuels that emit high levels of greenhouse gases [8]. Therefore, increasing importance is attached to the inclusion of the harvested wood products pool (wood products in use that continue to store carbon after harvest), avoided emissions, and LCA (life cycle assessment) indicators. Only such an expanded perspective allows the actual climatic effect of the forest-wood system to be estimated [5,9]. The digitisation of the wood supply chain and the development of tools for tracking material flows further increase the possibilities for precise carbon accounting at the operational level. This creates a basis for models that are more firmly grounded in management practice [10]. Current analyses of forests as carbon sinks must therefore include the site and stand component, the soil component, the product component, and the technological component. These elements form a configuration of variables that justifies the methodological logic adopted in the present research [7].

In response to the complexity of carbon balance analyses, machine learning methods are increasingly being used, allowing for precise modelling of nonlinear and multidimensional interactions in environmental data [11]. The literature has shown that ensemble models, boosting methods, and hybrid architectures can effectively estimate carbon stocks, sequestration, and forest biomass. They often outperform classical approaches in predictive accuracy and generalization capacity [12]. This is particularly important for variables that are difficult to measure directly or highly variable in space, such as soil organic carbon stocks. In such cases, the combination of field data, environmental information, and machine learning (ML) methods is now regarded as one of the most promising directions of

development [13,14]. Recent studies also indicate that the advantage of machine learning models should not be assessed solely through predictive performance. Resistance to overfitting, appropriate feature selection, elimination of data leakage, and interpretability of results are equally important [11,15]. Approaches combining predictive modelling with variable influence analysis are therefore of particular significance. This includes the use of SHAP tools (Shapley Additive Explanations) (a method for explaining the contribution of each predictor to an individual model output), which makes it possible to identify the direction and magnitude of the effect of individual features on model output and thus enhance the applied value of predictions in environmental management [11,16]. This justifies the selection of hybrid and ensemble algorithms in the present study, and the interpretation of the model not merely as a black box, but as a tool supporting the identification of mechanisms shaping gross CO₂e (carbon dioxide equivalent) in forest stands [11,16].

Under Polish and, more broadly, Central European conditions, the need to develop such models is particularly evident because the region is characterized by the simultaneous influence of climatic pressure, the major significance of the forest sector for the climate balance, and high site and management variability among stands [1]. From a methodological perspective, this means that tools must be developed that are capable of predicting the total carbon balance not on the basis of a single pool. Instead, they should use readily available field and site parameters, such as age, height, tree density, species composition, forest site type, and management variant, supplemented with information on soil, operational emissions, and environmental and circularity indicators [17]. Such an approach is important both cognitively, because it captures the multifactorial nature of the CO₂e balance (carbon dioxide equivalent), and practically, because it provides a basis for reducing costly direct measurements and for supporting decision-making in forest management, carbon inventory, and the assessment of management scenarios [4,17].

Although the final CO₂e balance in a circular economy (CE) depends on integrated processes across the supply chain and on the reuse of raw materials, the foundation of this system is a precise understanding of the initial biomass resources. For this reason the aim of the present study was to develop and optimize a predictive model for estimating the total baseline carbon balance, expressed as gross CO₂e (gross carbon dioxide equivalent), in Polish forest stands. The principal task was to produce a tool capable of accurately predicting carbon dioxide equivalent on the basis of measurable field-data characteristics, including age, height, the number of trees per hectare, and forest site type. This allows for a reliable estimate of the biological carbon stored in the forest, which serves as a starting point for further circular management without the need for costly and complex biomass measurements. Instead, it relies on more readily available parameters related to the tree stand and site conditions. The developed solution directly supports the objectives of the Forest Carbon Farms project (FCF) (Forest Carbon Farms), whose purpose is to create an advanced carbon inventory model adapted to the specific characteristics of Central European forests. The implementation of these research tasks is consistent with Poland's strategic commitments concerning the enhancement of the absorptive capacity of the forest sector within climate policy by 2030.

2. Materials and Methods

2.1. Research Design and Analytical Framework

The methodological framework adopted in this study was based on the joint analysis of site-related, stand-related, carbon-related, operational, and environmental variables to model the gross CO₂e balance of forest stands under circular-economy conditions. The modelling procedure assumed that gross CO₂e [Mg/ha] could be described using a heterogeneous set of predictors reflecting forest site type, stand composition and structure,

selected carbon pools, management-related emissions, and environmental indicators. This approach reflects the multifactorial nature of the forest carbon balance and supports the integrated interpretation of the ecological and management-related determinants of the analyzed system [18]. In this way, the analytical framework combined predictive and interpretative functions, linking model performance with the ecological and management-related structure of the analyzed forest system.

The adopted analytical framework was also consistent with the broader logic of forest assessment and monitoring, in which stand characteristics, site conditions, soil-related parameters, water-related factors, and selected environmental indicators are considered jointly. In this context, references to large-scale forest inventory and monitoring systems in Poland should be understood as a conceptual basis for the selection and integration of variables, rather than as a direct indication that the analytical dataset was derived exclusively from these systems [19,20]. This justified the construction of predictive models combining dendrometric, site, carbon, operational, and ecological variables within a single analytical procedure [10,21]. The same logic also supported the joint inclusion of stand structure, site conditions, carbon pools, management-related emissions, and environmental functioning within one coherent modelling scheme [22].

2.2. Research Material

The research material was treated as a curated analytical dataset prepared for predictive modelling at the forest stand or sample plot level, rather than as a probability-based representative sample in the sense of a national statistical forest inventory. The primary dataset contains information on forest management in Poland from 16 voivodeships and their macroregions (for the years 2020–2024). The dataset was prepared to enable the analysis of the effects of carbon sequestration and greenhouse gas emissions in different regions on forest management within a circular-economy framework [22]. The spatial scope included forest stands assigned to the analyzed macroregional conditions of Poland, with particular emphasis on Pomerania, Mazovia, and Greater Poland. In the present study, these regions should be understood as the principal macroregional reference framework for comparative and predictive analyses, rather than as a basis for claiming full statistical representativeness for all forests in Poland.

A single observation in the analytical dataset represented a forest stand defined by location, observation year, and management variant, with the main quantitative variables expressed per hectare. Only records meeting all of the following conditions were included in the analytical dataset: (i) assignment to a clearly defined observational unit at the forest stand or sample plot level; (ii) availability of a core set of habitat, stand, and emission-related variables; (iii) assignment of the record to a single year within the 2020–2024 period; and (iv) logical consistency of the record with the adopted structure of input variables used in the circular-economy-oriented forest carbon analysis. Incomplete, inconsistent, or target-deficient records were removed during preprocessing, resulting in 930 final records. Plot area was retained in the dataset as an auxiliary predictor rather than treated as the primary unit of analysis, ensuring consistency between the observational unit and indicators recalculated per hectare, but it was not treated as the basis for formal regional- or national-scale statistical inference. All records were handled as independent analytical observations at the stand or plot level, and the final modelling table was screened for duplicates and incomplete entries before model training.

The adopted dataset was not designed as a formally stratified sample representing the entirety of Polish forest resources. Therefore, the results should be interpreted primarily as applying to forest stands with habitat, structural, and management characteristics similar to those represented in the analyzed dataset. Generalization to all Polish forest macroregions

should be made with caution and would require additional validation using an independent and spatially stratified dataset.

The use of this type of research material was consistent with the aim of the study, which was not to reconstruct official national forest statistics, but to develop and validate a predictive tool capable of modelling gross CO₂e balance from stand-related, habitat-related, carbon-related, operational, and environmental variables considered jointly. In this sense, the dataset functioned as an operational modelling resource for identifying relationships between measurable forest descriptors and the gross carbon balance under circular-economy-oriented forest management. This way of organizing data aligns with the direction of Polish work on the carbon balance in forests, as the national carbon balance model is based on data concerning tree species traits and soils, and in a broader context also considers the long-term, persistent carbon sequestration in products.

The input variables were organized into five functional groups (Table 1). The first group consisted of habitat-stand variables, such as voivodeships, macroregions, the shortened forest habitat type, the dominant stand type or mixed arrangement, management variant, main species, bonitation, and sample surface area. The second group included biometric and structural variables, including trunk circumferences (L1, L2) and features related to the condition and structure of the stand: “Age years”, “Number of trees-estimated data [ha]”, “Height-estimated data [m]”, “Volume [m³/ha]”, “Dry mass [t/ha]”, “Deadwood [m³/ha]”, “Biomass left dry/ha”, “Share of remaining biomass [%]” (Table 1). The third group comprised carbon variables: above-ground biomass (AGB), below-ground biomass (BGB), the organic carbon stock in soil in the 0–30 cm layer (C SOC 0–30 Mg/ha), annual change in soil carbon stock (Δ SOC Mg/C ha per year), and carbon bound in products from harvested wood (C HWP Mg/C ha). The fourth group consisted of operational and circular variables, covering emissions from harvesting, transport, and processing, as well as emissions avoided through substitution of materials or fuels with high emissions. The fifth group included environmental variables (annual precipitation, annual temperature) and life cycle indicators, such as the water retention index, biodiversity index and soil compaction. All variables listed in Table 1 served as input data, and the aim was to identify a single optimal model.

Table 1. Characteristics of the initial set of input variables (field and operational descriptors) subjected to the optimization process.

Group	Variable Name	Variable Status
Habitat and stand variables	Shortened forest site type	Input
	Dominant tree stand type or mixed system	Input
	Management variant	Input
	Main species	Input
	Bonitation	Input
	Sample area [ha]—data from a forester	Input
	Macroregion	Input
	Voivodeship	Input
	Year of observation	Input
	Pine share [%]	Input
	Oak share [%]	Input
	Beech share [%]	Input

Table 1. *Cont.*

Group	Variable Name	Variable Status
Biometric and structural variables	Spruce share [%]	Input
	Fir share [%]	Input
	Alder share [%]	Input
	Age (years)	Input
	Number of trees—estimated data [ha]	Input
	Height—estimated data [m]	Input
	Volume [m ³ /ha]	Not used
	Dry mass [t/ha]	Not used
	Deadwood [m ³ /ha]	Input
	Biomass left dry/ha	Input
Carbon variables	Share of remaining biomass [%]	Input
	AGB [MgC/ha]	Not used
	BGB [MgC/ha]	Not used
	C SOC 0–30 MgC/ha	Input
	SOC MgC/ha/year	Input
	C HWP MgC/ha	Input
	Gross CO ₂ e t/ha	Output
	Net CO ₂ e t/ha	Not used
	CI95 t/ha	Not used
Operational and circular variables	Emissions from harvesting tCO ₂ e/ha	Input
	Emissions related to the transportation of wood or biomass tCO ₂ e/ha	Input
	Emissions from processing tCO ₂ e/ha	Input
	Avoided emissions tCO ₂ e/ha	Input
	Harvesting [m ³ /ha/year].	Input
Environmental and life cycle variables	Annual precipitation [mm]	Input
	Annual temperature [°C]	Input
	Water retention index	Input
	Biodiversity index of water retention	Input

The solution implemented allows for the assessment of the forest's condition, the main carbon pools, and the timber cycle in the context of a circular economy.

2.2.1. Habitat and Stand Variables

The first group consisted of habitat and stand variables, such as a shortened forest habitat type (“Shortened forest site type”), the dominant stand type or mixed arrangement (“Dominant tree stand type or mixed system”), management variant, main species, bonitation, and sample plot area (“Sample area [ha]—data from a forester”).

The shortened forest habitat type was determined as a synthetic description of habitat conditions, combining information on moisture, fertility, and the general character of the forest habitat. This variable corresponded to the forest typology used in forest management practice and was used to classify records into categories such as fresh pine (_Bśw),

fresh mixed pine (_BMśw), fresh mixed forest (_LMśw), fresh forest (_Lśw), or ash alder carr (_OIJ).

The dominant stand type or mixed arrangement was determined based on the species composition of the analyzed object. In the case of monospecific stands, the name derived from the dominant species was used, whereas for mixed stands, the main co-occurrence pattern of species was recorded, e.g., beech (_bukowy), oak ("_dębowy"), pine ("_sosnowy"), birch ("_brzozowy"), alder ("_olchowy"), pine-oak ("_sosnowo-dębowy"), pine-birch ("_sosnowo-brzozowy"), pine-beech ("_sosnowo-świerkowy"), pine-spruce, mixed coniferous-deciduous stands ("_mieszane iglasto-liściaste"), or others. This variable aimed to reflect a simplified species structure of the object and served as a feature describing both the production potential and functional diversity of the stand.

The management variant was defined as a scenario of forest management practices applied to the analyzed stand. In practice, it indicated the adopted model of biomass utilization and management, differentiating the level of intensity, the share of protective functions, and the scope of raw material use within a circular economy framework. This variable was assigned expert knowledge at the record level and used to distinguish more production-oriented, sustainable, protective, or circular variants.

The main species was identified as the dominant species in the analyzed stand, i.e., the one with the largest contribution to the upper layer or considered the most important from the perspective of the object's characterization. Abbreviations consistent with forestry practice were used, e.g., So for pine, Św for spruce, Db for oak, Bk for beech, Brz for birch, Ol for alder, and Jd for fir. This variable had biological and ecological significance, as the main species influences growth rate, resource availability, drought resistance, carbon sequestration processes, and the type of wood utilization possible.

Bonitation was defined as the productivity class of the habitat for the given main species. It did not refer to soil bonitation classes but to the quality of growth conditions for the stand. In practice, bonitation reflected the habitat's capacity to produce growth and was interpreted as an ordinal feature on a five-point scale (classes I, II, III, IV, and V), where a lower class number indicated better growth conditions. This variable was retained in the model as an ordered predictor. This ensured the proper application of SHAP analysis for model interpretation, allowing verification of how habitat conditions (classes I to V) influence the deviation of predictions from the baseline value.

Sample plot area was defined as a designated fragment of the stand serving as an observation unit, on which measurements and features of the object were taken or assigned. This variable allowed for converting absolute values to per-unit-area figures and maintaining consistency between measurement data and indicators related to 1 hectare. In forest monitoring logic, sample plots form the basic unit for collecting data on species, age, diameter at breast height, health status, and stand structure [23].

2.2.2. Biometric and Structural Variables

The second group consisted of biometric and structural variables, including trunk circumferences (L1, L2) and features related to the condition and structure of the stand. L1 was interpreted as the trunk circumference measured at a height of 1.30 m above ground level, which is equivalent to the traditional breast height recorded in circumference units. This variable described the basic dendrometric dimension of the tree and was treated as an indicator of the size of the individual or the average size of trees in a given record. L2 was interpreted as the trunk circumference at the base, measured at a low height above ground of about 0.05 m. Unlike L1, this variable better reflected the expansion of the trunk base and served as an auxiliary measure when describing the shape of the shoot, the relationship between breast height and the lower part of the trunk, and for intermediate differentiation

of species morphology. During data quality control, a logical condition was checked to ensure that L2 remained greater than L1. This group also included all features describing the condition and structure of the stand, such as age or age class, species composition, degree of structural diversity, crown condition, or overall health. Their selection was consistent with forest monitoring methodology, which records species, age, breast height, status, biosocial position, and numerous morphological features of crowns, including defoliation, discolouration, and the presence of dead branches, on observation plots.

2.2.3. Carbon Variables

The third group consisted of carbon variables: above-ground biomass (AGB), below-ground biomass (BGB), organic soil carbon stock in the 0–30 cm layer (C SOC 0–30 MgC/ha), annual change in soil carbon stock (Δ SOC MgC/ha/year), and carbon stored in products derived from harvested wood (C HWP MgC/ha).

AGB was determined as the carbon stock accumulated in the above-ground biomass of the stand, including trunk, bark, branches, and, depending on the adopted methodology, also leaves or needles. This variable was expressed in Mg C/ha, i.e., tonnes of carbon per hectare. In practice, the AGB value could be established based on biometric data and appropriate biomass–carbon conversion factors. This variable represented the primary pool of living above-ground biomass carbon. However, due to its direct relationship with the target variable, it was not retained in the final predictor matrix.

BGB was calculated as the carbon stock stored in below-ground biomass, primarily in the root system. Like AGB, it was expressed in Mg C/ha. BGB was regarded as a supplementary pool of living biomass carbon, particularly important for a complete carbon balance of the ecosystem. Due to its direct relationship with the target variable, BGB was not retained in the final predictor matrix.

C SOC 0–30 MgC/ha indicated the organic soil carbon stock in the 0–30 cm layer, expressed in tonnes of carbon per hectare. This variable characterized the size of the main carbon pool stored in mineral and humus-rich soil. Under forest conditions, it is one of the most important long-term carbon pools, and its inclusion was necessary for a proper assessment of the impact of forest management on the climate.

Δ SOC MgC/ha/year was defined as the annual change in organic soil carbon stock, i.e., the increase or decrease in SOC over the course of a year. Positive values indicated carbon accumulation in the soil, while negative values indicated loss. This variable played a special role in the model because it allowed separation of the stock's state from its dynamics.

C HWP MgC/ha represented the amount of carbon stored in products derived from harvested wood, converted to the unit of forest stand area. This variable reflected the fact that part of the carbon leaving the forest ecosystem remains bound in sawn timber, panels, furniture, structures, or other wood products, and does not immediately return to the atmosphere. Including this variable in the model was consistent with the national approach to developing methodologies for long-term, sustainable carbon sequestration in products. The Forest Condition Report explicitly states that work is underway in Poland on a methodology for long-term carbon sequestration in products and on a carbon balance model based on species and soil characteristics.

2.2.4. Operational and Circular Variables

The circular economy framework adopted in this study should be understood as a broader interpretative framework, whereas the operational variables represent its measurable process-related components. In the present analytical design, the circular economy dimension includes the retention of carbon in harvested wood products, cascading biomass use, and the estimation of avoided emissions resulting from material or energy substitu-

tion. By contrast, the operational variables refer specifically to greenhouse gas emissions generated during timber harvesting, transport, and industrial processing.

The fourth group consisted of operational and circular variables, including emissions from harvesting, transportation, and processing, as well as emissions avoided through the substitution of materials or fuels with high emissions. This distinction was introduced in order to separate the conceptual circular-economy context from the directly quantified emission-related predictors used in the models. To ensure methodological transparency, operational and circular variables were quantified within clearly defined system boundaries. The system included the following stages: (i) harvesting and timber extraction within the forest stand, (ii) transport of harvested biomass or timber to the first point of further utilization, and (iii) industrial processing of wood into intermediate or final products. Emissions were expressed in tonnes of CO₂ equivalent per hectare and were assigned to the stand-level analytical unit by relating process-specific emission factors to the biomass or timber output attributed to a given stand.

Emissions from harvesting were defined as greenhouse gas emissions generated at the stage of felling, pruning, cross-cutting, skidding, forwarding, and other operations related to timber extraction. They were expressed in tonnes of CO₂e per hectare. This variable represented the direct climate cost of forestry work associated with raw material utilization, and these emissions were quantified using fuel- or operation-based emission factors assigned to mechanized forestry activities.

Transport emissions referred to emissions assigned to the movement of wood or biomass from the point of harvest to the point of further use, such as a wood yard, sawmill, paper mill, or other processing plant. This value was also expressed in tonnes of CO₂e per hectare. It was linked to supply chain logistics and constituted the second main component of operational emissions. Emissions from transport were estimated using transport emission factors expressed per tonne-kilometre or cubic metre-kilometre, depending on the available conversion basis.

Processing emissions included emissions related to industrial wood processing, such as chipping, drying, production of panels, pulp, paper, or wood fuels. Their magnitude reflected the energy intensity of further processing of the raw material, and they were quantified using process-level emission factors linked to the relevant product pathway.

Avoided emissions were interpreted as the amount of emissions that were prevented thanks to the use of wood or biomass as a substitute for materials and fuels with higher emissions. This variable expressed the effect of material or energy substitution and was particularly important from a circular economy perspective. In practice, it reflected the climate benefit resulting from a wood product or circular solution replacing an alternative with a higher carbon footprint. Accordingly, avoided emissions did not represent direct measured emissions, but rather the climate benefit resulting from substitution. In each case, the avoided-emission term was calculated against an explicitly defined reference scenario and expressed in tonnes of CO₂ equivalent per hectare.

All variables in this group were treated as characteristics describing the operational dimension of wood utilization and its role within the circular economy logic. Their application was consistent with the fact that in Polish forests, a significant part of harvesting is not only related to planned utilization but also to the removal of dry, damaged, and fallen trees caused by climate stresses, water disturbances, and biotic factors. The analytical procedure therefore distinguished between (i) direct operational emissions, including harvesting, transport, and processing, and (ii) circular-economy benefits, including carbon retention in harvested wood products and avoided emissions due to substitution. This clarification was introduced to avoid conflating the circular economy as a system-level concept with the narrower group of operational predictors used in the machine learning models. All

emission-related variables were parameterised using fixed emission factors or scenario-based assumptions applied consistently across the dataset. Table 2 presents the emission factors, units, system boundaries, and data sources used to quantify these variables. Where a single direct field measurement was not available, the values were assigned using documented analytical assumptions consistent with the adopted modelling framework.

Table 2. Emission factors, system boundaries, and assumptions used for the quantification of operational and circular variables.

Variable	Definition	Unit	System Boundary
Emissions from harvesting	Emissions generated during felling, pruning, cross-cutting, skidding, and forwarding	tCO ₂ e/ha	Forest operations within stand boundary
Emissions from transport	Emissions related to timber or biomass transport from stand to first point of utilization	tCO ₂ e/ha	From stand to yard/sawmill/plant gate
Emissions from processing	Emissions generated during industrial wood transformation	tCO ₂ e/ha	Gate-to-gate or cradle-to-gate, depending on product pathway
Avoided emissions	Emissions avoided through substitution of high-emission materials or fuels	tCO ₂ e/ha	Relative to defined reference scenario
Carbon stored in harvested wood products	Carbon retained in wood products after harvest	MgC/ha	Post-harvest product pool
LCA GWP	Global warming potential per unit volume of raw material or product	kgCO ₂ e/m ³	According to adopted LCA scope

2.2.5. Environmental Variables and Life Cycle Indicators

The fifth group consisted of environmental variables and life cycle indicators, such as the water retention index, biodiversity index and soil compaction. The water retention index was treated as a synthetic indicator of the habitat or stand's capacity to retain water. It was most often recorded on an ordinal scale, e.g., from 1 to 5, where higher values indicated greater retention potential. This variable was intended to reflect the influence of habitat conditions, stand structure, and soil characteristics on water management within the forest ecosystem. Its inclusion was justified by monitoring data, which identified water conditions and soil moisture as some of the main factors affecting forest health [5,18].

The biodiversity index was defined as a synthetic indicator of the level of biological diversity within the stand or habitat, usually recorded on a scale from 0 to 1. Higher values indicated greater species or structural diversity. This variable was meant to represent the ecological dimension of ecosystem stability and was important both for emission prediction and for the later stage of multi-criteria optimization.

Soil compaction was interpreted as an indicator of the physical condition of the soil, relating to its degree of compaction or resistance to penetration. In practice, this variable could be expressed in MPa and used as an approximation of the degree of compaction affecting root growth, soil profile aeration, and water infiltration. In the model, it served as an auxiliary feature describing the physical conditions of stand functioning.

LCA GWP kgCO₂e per m³ represented the global warming potential expressed in a life cycle analysis per unit volume of raw material or wood product, i.e., in kilograms of CO₂ equivalent per cubic metre. This variable allowed linking traditional forest analysis with environmental assessment of the entire value chain, from timber harvesting to processing and further utilization. As a result, the model did not only refer to processes occurring

within the forest ecosystem itself but also partially accounted for the emissions associated with the material throughout its life cycle.

For the purposes of further modelling, the input variables were additionally treated as five complementary functional groups: site- and stand-related variables, biometric and structural variables, carbon-related variables, operational and circularity-related variables, and environmental variables. This division facilitated the subsequent interpretation of predictor importance in the machine learning models and supported the simultaneous inclusion of site characteristics, stand structure, carbon pools, and emissions associated with wood utilization within one coherent analytical dataset [11,21].

2.3. Data Preprocessing

In the first stage, it was necessary to clean the dataset concerning forest management in .csv format. This involved checking for duplicate records, missing values, and data format correctness. The original dataset contained 960 records. As a result of this procedure, the last 30 observations were removed because they did not contain complete values. The secondary dataset (df_cleaned) therefore contained 930 records. Each record represented aggregated data at the level of a single forest stand. The key values were standardized per hectare. This ensured consistency between the observation unit, the predictor variables, and the target variable used in the modelling procedure.

In the next step, a decision variable was defined. The target variable described the total carbon balance expressed as gross CO₂e [Mg/ha]. Operationally, the gross carbon dioxide equivalent balance was defined as the sum of the main carbon pools, calculated for the *i*-th analytical unit as follows (Equation (1)) [24]:

$$\text{Gross CO}_2\text{e}_i = \frac{44}{12}(\text{AGB}_i + \text{BGB}_i + \text{C}_{\text{soc},0-30,i} + \text{C}_{\text{HWP},i}) \quad (1)$$

where Gross CO₂e_{*i*} is the gross carbon dioxide equivalent balance for the *i*-th forest stand or sample plot [t CO₂e ha^{−1}], AGB_{*i*} is the carbon stock accumulated in above-ground biomass [Mg C ha^{−1}], BGB_{*i*} is the carbon stock accumulated in below-ground biomass [Mg C ha^{−1}], C_{SOC,0–30,*i*} is the stock of soil organic carbon in the 0–30 cm layer [Mg C ha^{−1}], and C_{HWP,*i*} is the carbon stored in harvested wood products [Mg C ha^{−1}]. The factor 44/12 converts carbon stock into carbon dioxide equivalent on the basis of molecular mass.

For interpretative purposes, the gross balance was further related to an extended net-balance formulation that additionally included annual soil carbon dynamics, operational emissions, and avoided emissions associated with substitution effects (Equation (2)) [24]:

$$\text{Net CO}_2\text{e}_i = \text{Gross CO}_2\text{e}_i + \frac{44}{12}\Delta\text{SOC}_i - (\text{E}_{\text{harv},i} + \text{E}_{\text{trans},i} + \text{E}_{\text{proc},i}) + \text{E}_{\text{avoid},i} \quad (2)$$

Gross CO₂e was available in the dataset as a separate response variable and was not recalculated from the predictors used in the final models. Because the target variable used in machine learning was the gross CO₂e balance (Equation (1)), special attention was given to avoiding circularity between the response variable and the predictors. Variables that were directly or algebraically related to Equation 1, particularly AGB, BGB, dry biomass mass, and net CO₂e (Equation (2)), were excluded from final predictor matrix before model training. Variables related to harvested wood products, avoided emissions, soil organic carbon change, and operational emissions were retained only as descriptive or contextual variables to support the interpretation of the broader carbon balance framework, provided they did not introduce direct circular dependence. Therefore, the final machine learning models were based on independent stand, site, structural, management, and environmental descriptors rather than on direct algebraic components of the response variable.

In the subsequent stage, it was necessary to ensure the correctness of textual data by standardizing the numeric format. The decimal separator was replaced from a comma to a period, enabling an attempt to convert these columns into numeric data. Then, all text columns were converted to a floating-point data type. Columns containing actual descriptive text were left in string format. Finally, the decision variable was converted to a floating-point type, which prepared the data for further statistical analysis and modelling.

Another important step in data preparation was the practical implementation of the variable exclusion procedure described above. During this stage, four variables were removed from the dataset: AGB (aboveground biomass), BGB (belowground biomass), dry biomass mass, and CO₂ equivalent per ton per hectare in a net perspective (Table 1). These variables were excluded because they were directly related to, or derived from, the target variable and could therefore introduce circularity into the modelling procedure. This step reduced the risk of target leakage and artificially inflated model performance. Additionally, correlation analysis was conducted to identify predictors that were too strongly correlated with each other or with the response variable. Identifying and reducing such redundant relationships was important for optimizing the input data and limiting model instability. As a result of this analysis, a very strong correlation was observed between “CI95 t/ha” and “Volume [m³/ha]”. This finding justified reducing redundant information in the predictor set and supported a more cautious interpretation of the high predictive performance of the models.

The next step was to divide the dataset into training, validation, and test sets. To extract data for the validation set, the data split was carried out in two stages. In the first stage, the dataset was randomly divided into a training set and a validation/test set in an 70:30 ratio. The second stage involved dividing the validation and test sets in a 50:50 ratio. As a result, the dataset was split into 70:15:15 proportions for the training, validation, and test sets, respectively. At this stage, the `train_test_split` function from the `model_selection` module of `scikit-learn` was used, with the above proportions specified via the `test-size` parameter. To ensure reproducibility of the results, the `random-state` parameter was also considered.

Next, a mini-pipeline was created to process numerical features in the dataset. To do this, data fitting was performed on the training set. The `SimpleImputer` function with the `strategy` parameter was used, aiming to fill in missing NaN values with the median of the respective column [25]. This method provided robustness against outliers and prevented missing data in the training set. The use of the median in `SimpleImputer` was justified, among other reasons, by the data on organic carbon stock (SOC), which varied significantly depending on soil type and region (e.g., pine forests in northern Poland and beech stands in the central part of the country), generating natural extreme values across the entire dataset. Besides the `SimpleImputer` method, the pipeline also included the `StandardScaler`, which aimed to set the mean close to 0 and the standard deviation equal to 1 for the independent variables. Incorporating these steps into the pipeline ensured they were performed in the correct order both during model training and when applying it to new data, guaranteeing consistency and eliminating the need for manual repetition.

In order to process textual data, a pipeline using the `OneHotEncoder` method was also implemented, transforming each categorical feature into binary columns [26]. Two arguments were used in the pipeline: `handle_unknown = ignore` and `sparse_output = False`. The first argument specified whether there was a risk of encountering categories in new test data that were not present in the training set. The second argument returned the encoded variables as a dense array, which ensured compatibility with the subsequent modelling steps.

In the data preparation process, the `ColumnTransformer` method was also used, enabling the transformation of input data into the appropriate feature space while considering

separate processing of numerical and categorical variables [27]. This approach allowed for clear separation of different data types, ensured reproducibility of processing operations, and reduced the risk of data leakage in both the training and test sets. All preprocessing operations were fitted exclusively on the training data and subsequently applied to the validation and test subsets, thereby preserving methodological consistency and limiting the risk of information leakage.

In order to optimize the model and reduce the dimensionality of the feature space, a variable selection method based on Lasso regression (L1 regularization) was applied using the `SelectFromModel` module [28]. This approach allowed for the automatic identification of the most important input variables and the elimination of those with marginal significance by setting their coefficients to zero. It simplified the model, significantly reducing the risk of overfitting. The entire procedure was integrated into a single processing pipeline, combining the initial data transformation step with an automatic feature selection process (Figure 1). This was especially important when analyzing the forest dataset, which contained numerous dendrometric, soil, and climatic parameters.

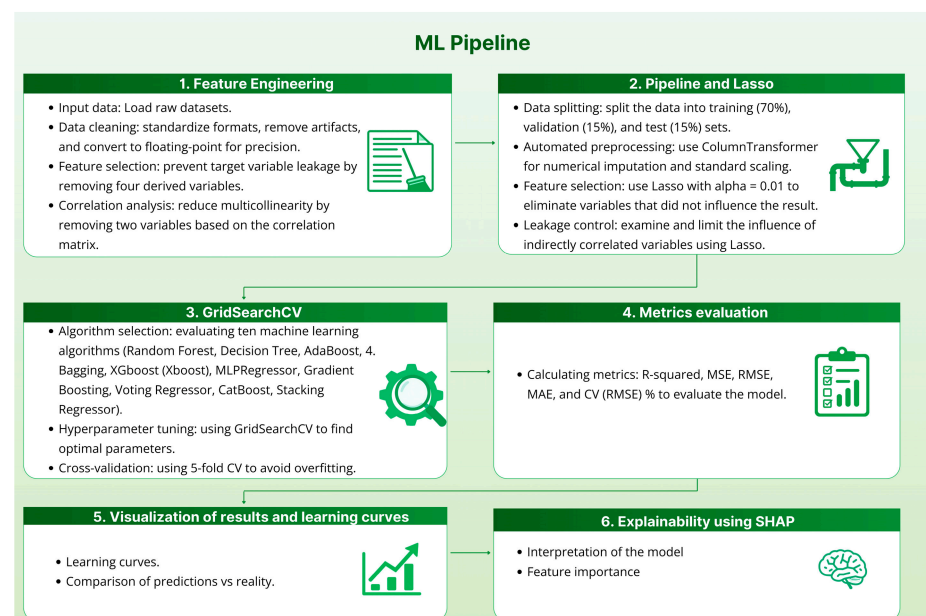


Figure 1. Flowchart of the ML process from data to interpretation.

2.4. Development of Machine Learning Algorithms

2.4.1. Neural Network and Hybrid Models

MLPRegressor is a neural network composed of an input layer, an output layer, and at least one hidden layer. The training process involved mapping complex, nonlinear relationships in the data by adjusting weights using selected learning optimizers [29–31]. The model's performance depended on the precise selection of chosen hyperparameters. The hyperparameter selection for this model was carried out as follows: `hidden_layer_sizes`, `activation`, `solver`, `alpha`, and `learning_rate_init` (Table 3). The `hidden_layer_sizes` parameter defined the network architecture by selecting the number of neurons in the hidden layers. Using the rectified linear unit (ReLU) activation function introduced necessary nonlinearity into the calculations, enabling the algorithm to better fit the patterns in the training set [32]. The “solver” parameter, responsible for the weight optimization algorithm, with “adam” valued for high efficiency [33], and “sgd” for stochastic gradient descent, ensured stability during the training process [34]. The regularization parameter `alpha` mitigated overfitting by penalizing excessively high weight values, improving the model's ability to generalize to new data. The hyperparameter `learning_rate_init` controlled the initial step size in the

weight update process, providing stability in learning and the rate at which the loss function reaches its minimum. As part of minimizing the loss function, i.e., the mean squared error (MSE), this model can achieve very high prediction accuracy in engineering applications.

Table 3. Hyperparameter configuration used in this research.

Model Name	Hyperparameter	Tested Values
Random Forest	max-depth	5, 6, 7
	n_estimators	200, 500
	max_features	sqrt, log2, 0.3, 0.5
	min_samples_leaf	1, 2, 4, 8
	min_samples_split	2, 5, 10
	bootstrap	True, False
Decision Tree	max-depth	3, 5, 10, 20
	min_samples_leaf	2, 5, 10
	min_samples_split	1, 2, 4
	ccp_alpha	0.0, 0.01, 0.5
AdaBoost	n_estimators	100, 200, 500
	learning_rate	0.01, 0.05, 0.1, 0.5, 1.0
	loss	linear, square, exponential
	estimator	DecisionTreeRegressor(max_depth = 3), DecisionTreeRegressor(max_depth = 5), DecisionTreeRegressor(max_depth = 7)
Bagging	n_estimators	50, 100, 200, 500
	max_samples	0.5, 0.7, 0.8, 1.0
	max_features	0.5, 0.7, 0.8, 1.0
	bootstrap	True, False
	bootstrap_features	True, False
	estimator	DecisionTreeRegressor(max_depth = 5), DecisionTreeRegressor(max_depth = 10)
XGboost	n_estimators	100, 200, 500
	max_depth	3, 5, 7
	learning_rate	0.01, 0.05, 0.1
	subsample	0.7, 0.8, 1.0
	colsample_bytree	0.7, 0.8, 1.0
	min_child_weight	1, 3, 5, 10
	reg_alpha	0, 0.01, 0.1, 1.0
	reg_lambda	0.5, 1.0, 2.0, 5.0
	gamma	0, 0.1, 0.5, 1.0
	hidden_layer_sizes	(50,),(100,)
MLPRegressor	activation	relu
	solver	sgd, adam
	alpha	0.001, 0.01
	learning_rate_init	0.00005, 0.0001, 0.001
	n_estimators	100, 200, 500
Gradient Boosting	max_depth	3, 4, 5
	learning_rate	0.01, 0.05, 0.1

Table 3. *Cont.*

Model Name	Hyperparameter	Tested Values
	subsample	0.7, 0.8, 1.0
	n_iter_no_change	10, 20, 50
Voting Regressor	estimators	reg1 = RandomForestRegressor(n_estimators = 100, random_state = 42) reg2 = BaggingRegressor(n_estimators = 100, random_state = 42) reg3 = xgb.XGBRegressor(random_state = 42)
CatBoost	iterations	500
	depth	4, 6, 8
	learning_rate	0.01, 0.05, 0.1
	l2_leaf_reg	1, 3, 5
Stacking Regressor	rf_n_estimators	100, 200
	rf_max_depth	3, 5
	xgb_max_depth	3, 5
	xgb_learning_rate	0.05, 0.1

Voting Regressor is the equivalent of a meta-learner in regression tasks. It is an advanced ensemble learning technique that integrates the predictions of several independent base models to generate a final, more stable result. This mechanism relies on aggregating the outputs of individual algorithms, which allows for effective reduction in variance and errors specific to single models, thereby increasing the algorithm's ability to generalize to new data [35,36]. In the research, a combination of Random Forest, Bagging, and XGBoost was used. Recent studies have shown that hybrid solutions often outperform individual neural network models in effectiveness. The selection of base models within the Voting Regressor structure has become particularly resistant to noise and overfitting in the research problem [37,38].

Stacking Regressor, like the Voting Regressor, integrates predictions from multiple different base models to generate a final, stronger result in the idea of a meta-model [39–41]. The procedure for using base models in this study was as follows: the first base estimator used was Random Forest (RF). RF aimed, through its structure of multiple decision trees, to improve stability and reduce the variance of the model. The second estimator used was XGBoost (XGB), which aimed to prevent overfitting through efficient gradient boosting implementation, accuracy in modelling complex, nonlinear relationships, and built-in regularization mechanisms. The parameter `n_estimators = 100` was used for the random forest, and the `reg:squarederror` loss function for XGBoost, providing a standard, stable configuration for regression tasks. The combination of these two algorithms allowed for mutual complementarity between Random Forest and XGBoost, by eliminating noise and precisely correcting forecasting errors made in previous stages (Table 3).

2.4.2. Advanced Gradient Models

Gradient Boosting Regressor is an ensemble algorithm designed to create successive decision trees by gradually correcting errors in previous models [42–44]. The result of this approach is improved forecast accuracy. The hyperparameter configuration of the model was carried out according to the scheme below. The parameter `n_estimators` was chosen to allow the model to account for more complex dependencies among the trees in the ensemble. The `max_depth` parameter limited the depth of individual trees, controlling the

complexity of each base model and the risk of overfitting. Regarding the `learning_rate`, the contribution of each new tree was scaled, with smaller values requiring a larger number of trees. The `subsample` parameter determined what fraction of samples was used for training each tree, introducing randomness and acting as a form of regularization. Meanwhile, `n_iter_no_change` was responsible for early stopping of the training process when the validation score did not improve after a specified number of iterations. In this algorithm, an early stopping mechanism (`n_iter_no_change`) was employed, which monitored the model's error based on predictions from internal validation subsets (out-of-fold predictions), ensuring that the final test set remained entirely isolated and was used solely for a reliable assessment of generalization [39].

XGBoost (Extreme Gradient Boosting) is a highly efficient implementation of gradient boosting that optimizes prediction by sequentially correcting errors of previous models [45–49]. This algorithm is based on a level-wise tree growth strategy, which ensures the stability of the tree structure while aiming to minimize a complex loss function. A distinguishing feature of this model is the built-in L1 (Lasso) and L2 (Ridge) regularization, which penalizes excessive complexity, effectively protecting the algorithm from overfitting. In the study, besides parameters such as `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, the stabilization parameter `min_child_weight` was selected, which determines the minimum sum of observation weights required in a child node (Table 3). An L1 regularization parameter was applied, promoting sparsity of weights and implicitly enforcing feature selection. Conversely, the use of L2 regularization penalizes large weight values, preventing model instability. The `gamma` parameter defined as the minimum reduction in loss required to split a node, was used as an additional mechanism for pruning the tree structure to simplify the model (Table 3). In research on the inversion of dye concentrations in wood based on hyperspectral data, this algorithm achieved excellent results, obtaining an R^2 value in cross-validation ranging from 0.928 to 0.947 [50].

CatBoost is an advanced gradient boosting algorithm, whose important feature is native and highly efficient processing of categorical features without the need for prior manual encoding. The use of innovative ordered boosting technique and a unique structure of symmetric decision trees effectively minimizes the risk of overfitting and offers exceptional speed during the inference process [51–53]. The machine learning algorithm is characterized by very high prediction accuracy and stability of results even when using default hyperparameter settings, making it an excellent tool for modelling complex dependencies in heterogeneous tabular datasets. In the studies, hyperparameters related to iterations, depth, learning rate, and `l2_leaf_reg` were used (Table 3). The iteration parameter defined the number of trees built during the training process (number of boosting rounds). In crop yield studies, values ranged from 67 to 91, while in wood colouring recipe prediction, this number was 300. The use of the `l2_leaf_reg` parameter as an L2 regularization coefficient (Ridge type) imposed a penalty for excessively high weight values in the tree leaves. In studies on wood properties, it was observed that after input data optimization, the value of this `l2_leaf_reg` parameter increased from 1 to 9, which allowed for a more stable and accurate model.

2.4.3. Tree-Based and Forest-Based Models

Random Forest is an advanced machine learning method based on an ensemble of models, utilizing a structure composed of multiple decision trees to solve complex classification and regression problems. Each tree within this ensemble adapts to a different subset of data, significantly increasing the algorithm's efficiency while reducing the risk of overfitting. In this research, UAV LiDAR data were used to determine individual tree height and crown radius, enabling the development of equations to estimate above-ground

biomass (AGB). This approach provided very high accuracy, with $R^2 = 0.9403$ [4]. In the current research, the `max_depth` parameter controlled the depth of the trees, and appropriate limiting of branching levels protected the model from overfitting [50,54]. Increasing the number of estimators in the `n_estimators` parameter improved the algorithm's ability to recognize complex physicochemical patterns in the data. The important `max_features` parameter ensured forest diversity by randomly selecting a subset of features at each node split. Regularization through `min_samples_leaf` and `min_samples_split` smoothed the tree structure and limited the influence of informational noise. The bootstrap method (bagging aggregation) reduced variance, leading to better generalization during attribute analysis in the dataset (Table 3).

A decision tree (DT) is a “glass box” algorithm, valued in engineering for its high interpretability and transparency of decisions [55–57]. This model operates by sequentially dividing the data set into smaller regions based on specific conditions, which allows for effective classification or regression of product parameters. In research estimating forest carbon stock (CS), decision tree-based ensemble models showed equally high performance when trained on the main dataset. The Bagged Decision Trees (BaggedDT) and Gradient Boosted Decision Trees (GBDT) algorithms achieved R^2 values of 0.90 and 0.93, respectively [12]. In the current studies, the hyperparameter grid again included, as in Random Forest, parameters such as `max_depth`, `min_samples_split`, and `min_samples_leaf` (Table 3). The `ccp_alpha` parameter was also used, supporting the tree's regularization process. This helped simplify its structure and increase its generalization ability. Optimized DT models are ideal for industrial applications because they allow for the direct translation of mathematical rules into technological instructions.

AdaBoost (Adaptive Boosting) is a traditional ensemble learning method that generates a strong predictive model by sequentially integrating multiple weak estimators, most often in the form of simple decision trees. An important mechanism of the algorithm is the dynamic updating of weights, which involves assigning greater importance to observations that were misclassified in previous stages, forcing subsequent models to correct these errors [58,59].

Bagging (Bootstrap Aggregating) is a machine learning ensemble technique that involves simultaneously generating multiple independent copies of the same classifier trained on different subsets of data. The main aim of this algorithm is to reduce the model's variance, which effectively minimizes the risk of overfitting and allows for better handling of informational noise in the training set [58].

2.5. The Optimization and Validation Process

In order to optimize each of the listed models (except for the Voting Regressor), hyperparameter grids (`param_grid`) were defined. The process of selecting optimal parameters was carried out using the GridSearch method, which systematically searches through defined parameter combinations to find those that minimize the model error [59]. For the purpose of ensuring high generalization ability, 5-fold cross-validation was also applied, which minimized the risk of overfitting through rotational data splitting [60,61]. Negative mean squared error (`neg_MSE`) was used as the evaluation metric, enabling a precise comparison of model quality across different hyperparameter combinations. Integrating the process into a processing pipeline eliminated the risk of data leakage, ensuring that feature transformations are estimated exclusively on the training subsets of each fold. This approach ensured a reliable and objective evaluation of the algorithms' performance on unseen data. Finally, the models were trained on the available research data, which allowed for the highest prediction accuracy of total carbon balance (expressed as gross CO_2e) for the studied stands.

2.6. Evaluation Criteria

As part of the assessment of the quality of regression models, the coefficient of determination (R^2) was used to account for the degree of variability of the dependent variable, the total carbon balance (expressed as gross CO_2e) [62,63]. This indicator typically ranges from 0 to 1, with a value close to 1 indicating a very good fit of the model's predictions to the actual results. This metric allows for direct comparison of model results across different datasets with varying scales of values without the need for additional transformations [62]. However, caution should be exercised to avoid overfitting, as the raw R^2 value always increases with the addition of new variables. Literature emphasizes that R^2 increases with the addition of new variables, which can lead to overfitting the model [64]. This issue is often mitigated through validation procedures, particularly cross-validation or evaluation on an independent test set. In this study, various feature selection methods, optimization techniques, and cross-validation were employed to address this.

Mean Squared Error (MSE) is a measure used in regression models, defined as the arithmetic mean of the squares of the differences between actual and predicted values. This measure heavily penalizes large deviations, making it very sensitive to outliers. In machine learning, MSE is commonly used as a loss function because its differentiability facilitates the mathematical optimization of model parameters during training. To obtain a more understandable measure of error, its square root is often taken, which is called RMSE [65,66].

Mean Absolute Error (MAE) is a regression measure calculated as the arithmetic mean of the absolute differences between the actual values and the model's predictions. Unlike MSE, this measure treats errors in a linear manner, making it much more resistant to the negative impact of outliers in the dataset. The main advantage of MAE is its easy interpretability, as it expresses the magnitude of the error in the same physical units as the predicted target variable. By applying the absolute value, this metric effectively prevents positive and negative errors from cancelling each other out, providing reliable information about the actual average deviation. In research, low MAE values indicate very high precision and satisfactory fitting of neural networks to experimental data [56].

2.7. Explainability of Variables Using SHAP

The SHAP (SHapley Additive exPlanations) tool is an advanced explainable artificial intelligence (XAI) tool that enables precise interpretation and understanding of how complex mathematical models make decisions. This technology provides consistent explanations at both the global and local levels, indicating which specific input variables had the greatest impact on the obtained gross carbon footprint per unit area [59,67–69]. In the present study, SHAP analysis was used for the interpretation of the final predictive models in order to identify the relative contribution of individual predictors to the estimated gross CO_2e values and to complement the purely accuracy-based comparison of algorithms.

2.8. Software

Numerical experiments were conducted on a Linux system based on the x86_64 architecture, which allowed for effective memory management and efficient execution of complex matrix operations. The calculations were performed using a stable version of the Python 3.12.13 interpreter. The entire environment was run on the Google Colaboratory cloud platform, providing access to hardware accelerators (GPU/TPU). Statistical analysis along with modelling was carried out in the Google cloud environment using the Python programming language, employing libraries such as Pandas, Numpy, Scikit-learn, xgboost, catboost, shap, matplotlib, seaborn, os, gc, conducting statistical tests, regressions, and machine learning algorithms. Canva's AI features (Canva Pty Ltd., Sydney, Australia)

were used for the preparation of graphical materials such as Figure 1. Grammarly (Grammarly Inc., San Francisco, CA, USA) was additionally utilized solely for language and grammar editing.

3. Results and Discussion

3.1. Model Evaluation Metrics

In the context of the research on estimating a predictive model for the total carbon balance (expressed as gross CO₂e), the analysis of performance indicators collected in the table allowed for an adequate assessment of the algorithms' ability to generalize environmental data contained in the dataset. Table 4 contains 10 machine learning models, which after reduction of highly correlated variables showed a good fit to the actual equilibrium. In the first training stage on the test set, 8 out of 10 models had an R² value above 0.99 before eliminating correlated variables. Ultimately, to prevent overfitting, the algorithms were re-trained by removing selected variables and changing the data proportions from 80:10:10 to 70:15:15, which resulted in 6 out of 10 models achieving R² > 0.90. It should be noted that the study used a random split of spatial and temporal forest data. According to the conclusions of Wadoux et al. [70], such an approach can yield overly optimistic results, especially with grouped data. This is due to the natural similarity of neighboring plots and the correlation across subsequent years. For this reason, the high values of the coefficient of determination (R² > 0.90) obtained may be overestimated, and caution should be exercised when interpreting them in the context of predictions for entirely new research areas or future years. To confirm the reliability of the results in both cases, the Lasso feature selection method and cross-validation were used to avoid overfitting. The final results of the second training stage are presented in Table 4. The results below show the final model configuration after the optimization process. In the current research on the CO₂e balance, the Decision Tree model achieved an R² of 0.8056, making it one of the four models with a lower R² than the others. Other algorithms, such as Bagging, Random Forest, and Voting Regressor, achieved results closer to R² = 0.9. The literature clearly indicates that ensemble models, such as XGBoost and CatBoost, attain the highest accuracy in modelling complex nonlinear dependencies [51,52,65]. XGBoost demonstrated excellent generalization ability in assessing tomato quality (R² = 0.98) [71]. Meanwhile, CatBoost achieved an R² of 0.95 for various colouring systems when modelling wood dyeing recipes [39]. In the current research on estimating the total carbon balance (gross CO₂e), MLPRegressor ranked second, and XGBoost fourth among the other algorithms. The meta-model from the stacking group, which combines base models like Random Forest and XGBoost, also achieved a high R² value, placing third among the models. Looking at individual algorithms, the Catboost model stands out, achieving the highest R² value of 0.9481.

The study also built a Voting Regressor metamodel as a hybrid model, which integrated the predictions of three independent base estimators: Random Forest, Bagging, and XGBoost. As a result of the vote (results) aggregation between the models, the metamodel appeared to be more resistant to specific errors than individual algorithms. Consequently, an R² value of 0.8923 made this algorithm seventh among a group.

Table 4. Performance metrics for the ultimate, optimized predictive models after the variable reduction process (test set).

Model	R-Squared (R ²)	MSE	MAE	RMSE	CV(RMSE)
Random Forest	0.892	48,047.770	126.606	219.198	29.767
Decision Tree	0.806	86,737.838	184.000	294.513	39.994
AdaBoost	0.912	39,114.962	116.812	197.775	26.857
Bagging	0.880	53,480.906	123.644	231.259	31.405

Table 4. Cont.

Model	R-Squared (R^2)	MSE	MAE	RMSE	CV(RMSE)
XGboost	0.933	29,985.796	100.751	173.164	23.515
MLPRegressor	0.940	26,721.980	104.901	163.469	22.199
Gradient Boosting	0.936	28,713.472	96.430	169.451	23.011
Voting Regressor	0.894	47,486.900	114.352	217.915	29.592
CatBoost	0.948	23,177.672	91.913	152.242	20.674
Stacking Regressor	0.934	29,462.084	105.886	171.645	23.309

In the comparative analysis, the results from Table 4 were examined based on the Root Mean Squared Error (RMSE) indicator, which is the square root of MSE. Similar to MSE, a lower RMSE value indicated a better model fit. To illustrate the relationship between the prediction quality indicators, a heatmap of these indicators was used, where the colours for RMSE reflected the same trends as for MSE. Analyzing the observed results on the test set, it can be seen that the lowest RMSE value was achieved by the Catboost model. Similarly, models with the smallest predictive errors on unseen test data included MLPRegressor, XGBoost, Gradient Boosting and Stacking Regressor, demonstrating better model fit. AdaBoost, Bagging, and Voting Regressor also achieved acceptable MSE and RMSE values, with RMSE ranging from 197.775 to 231.260. It is explained that a low RMSE value combined with a high coefficient of determination (R^2) indicates an exceptionally high effectiveness of the model fit to the data in this set. In the heatmap (Figure 2a–c), Decision Tree, Random Forest, and the Bagging algorithms were highlighted in darker colours, which correspond to their poorer model performance as measured by the MSE metric (Figure 2a). Furthermore, when analyzing the validation set (Figure 2b), it can be observed that Voting Regressor and AdaBoost also fall into this lower-performance group. It is worth noting that an analysis of the training set shows that four models perform worse, as directly reflected by the darker colours on the heat map using the MSE metric (Figure 2c). The heatmap visually clarified that 4 out of 10 models with the highest R^2 coefficient are noticeably lighter in the MSE and RMSE columns for the test set, indicating their better performance (Figure 2). Their lower MSE and RMSE values suggest that their predictions are closer to the actual values, and the models themselves generalize better to unseen data. Decision Tree, Random Forest and Bagging stand out with darker colours, which correlates with their poorer model performance results.

In order to facilitate the interpretation and comparison of results, the coefficient of variation in the CV(RMSE) was also calculated for the models on the test set [72–76]. The literature increasingly highlights that an excessive focus on a model's ideal accuracy (e.g., $R^2 > 0.99$), without considering its actual reliability, leads to models that perform well on training data but fail completely in real-world conditions due to overfitting [77]. Models exhibiting an unnaturally perfect fit with $R^2 > 0.99$ were rejected after the first training stage, as such high accuracy measures often do not translate into reliability and are merely the result of the algorithm's overfitting to the training sample. In the second training stage, after reducing highly correlated variables, a stable CatBoost model with a cross-validation error (RMSE) of approximately 20% was accepted, reflecting the natural variability of the studied parameters in a safer and more reliable manner. The Catboost model, achieved a CV(RMSE) value of 20.67% (Table 4). This meant that the prediction error was on average only about 20.5% relative to the mean value of the dependent variable (Table 4).

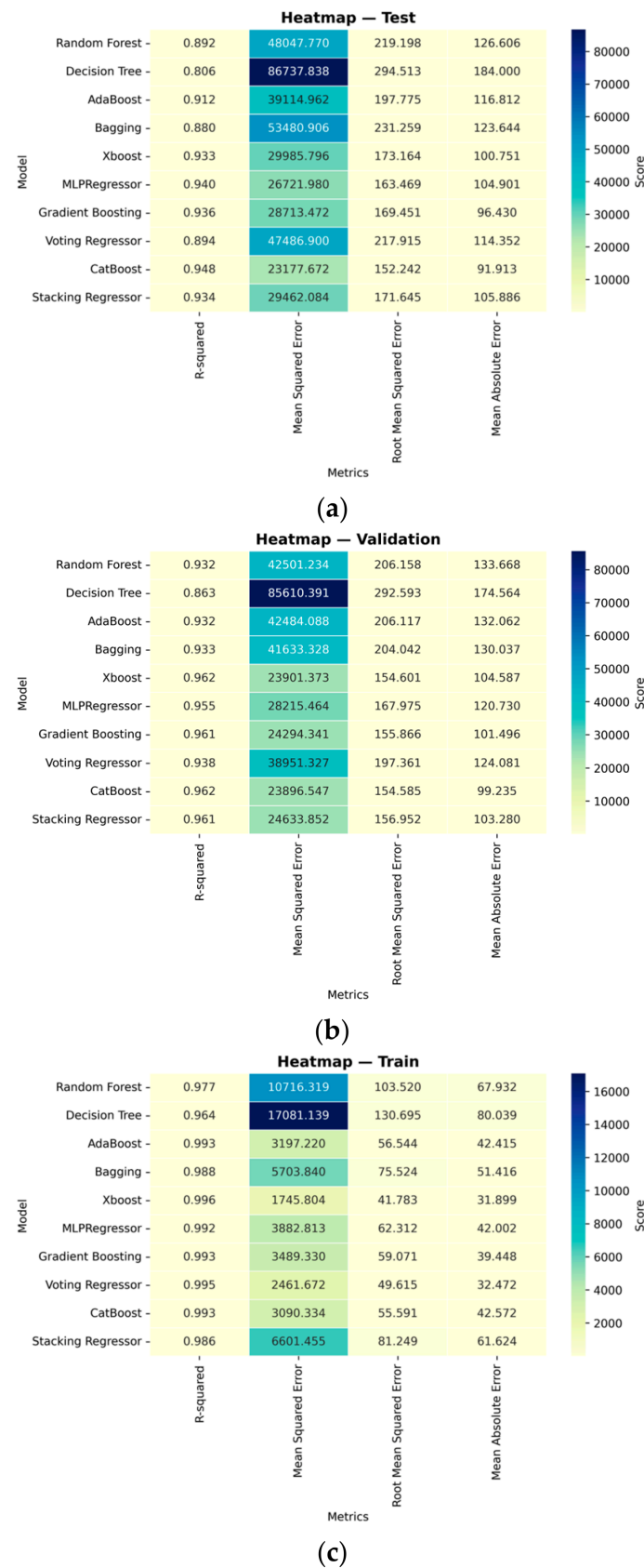


Figure 2. (a) Heatmap of performance metrics across different machine learning models (test set). (b) Heatmap of performance metrics across different machine learning models (validation set). (c) Heatmap of performance metrics across different machine learning models (training set).

3.2. Actual and Predicted Values

The presented chart determines the coefficient of determination (R^2), which indicates how well the model reflects reality by specifying the percentage of variance in the measured values (Actual) explained by the model's predictions (Predicted). The quality of the prediction is assessed against the ideal line (red dotted line), on which all points would lie in the case of completely error-free forecasting. The scatter of points around this axis illustrates the magnitude and nature of the error (blue points). Symmetry of the distribution suggests its randomness, while systematic deviations indicate bias in the model.

A comparative analysis of actual and predicted values allowed for a visual assessment of the models' accuracy by observing the degree of concentration of training cases relative to the line of perfect fit (Figure 3a–j). This phenomenon is confirmed in the literature, where, for example, the CatBoost algorithm used for predicting rice and corn yields showed that data points closely follow the theoretical regression line, resulting in high R^2 coefficients of around 0.90 [78]. In the current study, CatBoost achieved the highest predictive performance among all the methods tested, outperforming the other models in both accuracy and generalization (Figure 3g). Conversely, the weakest performance in terms of convergence with the fit line was observed for the Decision Tree model (Figure 3h), which is reflected in the highest coefficient of variation CV(RMSE) at 39.99%, as well as R^2 (Table 4). This result indicates that the prediction error of this algorithm is over 19% larger relative to the mean value of the studied decision variable than in other models. The poor fit of the decision tree model may suggest its lower stability or tendency to oversimplify the data structure compared to advanced neural networks or hybrid models. The figure shows a lack of point clustering around the perfect fit line for the Decision Tree. Significant deviations from the fit line are visible, which in practice could lead to incorrect estimates of the carbon (CO_2e) balance and reduce the reliability of this model. It is worth noting that the clear superiority of the CatBoost algorithm over a single Decision Tree (Figure 3h) confirmed that in tasks with high non-linearity, models based on the architecture of symmetric trees and the technique of ordered boosting provide higher accuracy and better generalization of patterns from the data [50]. It was observed that on the actual versus predicted value chart, where the CatBoost model's points are closest to the line of perfect fit (Figure 3g). This indicates less predictive noise and fewer extreme errors, known as outliers [50,79].

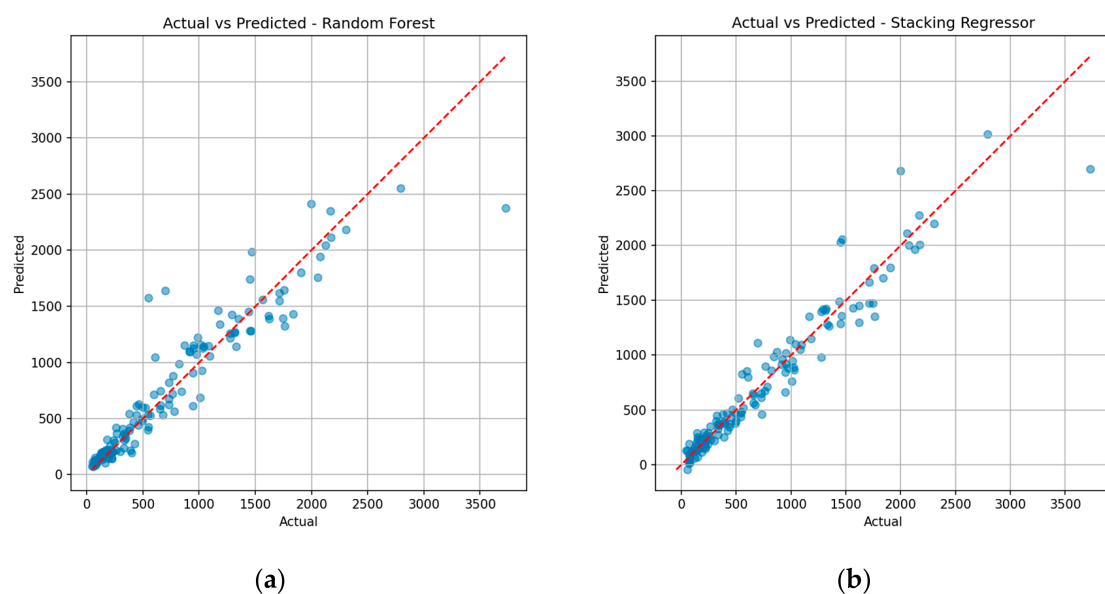
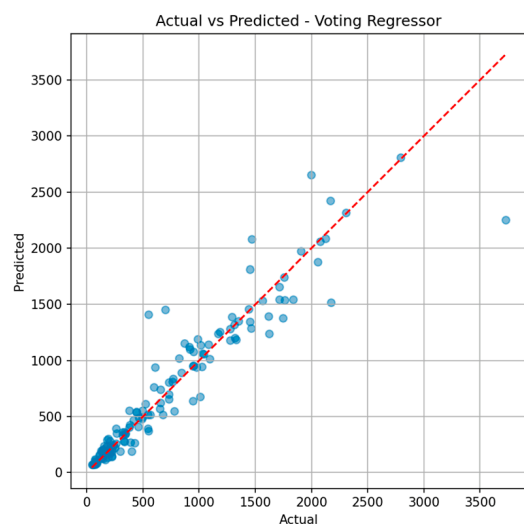
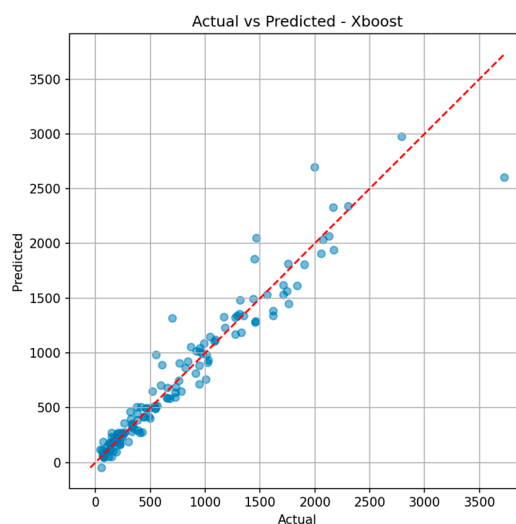


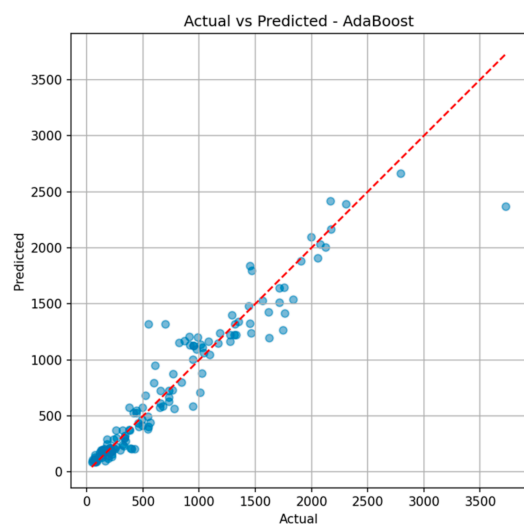
Figure 3. Cont.



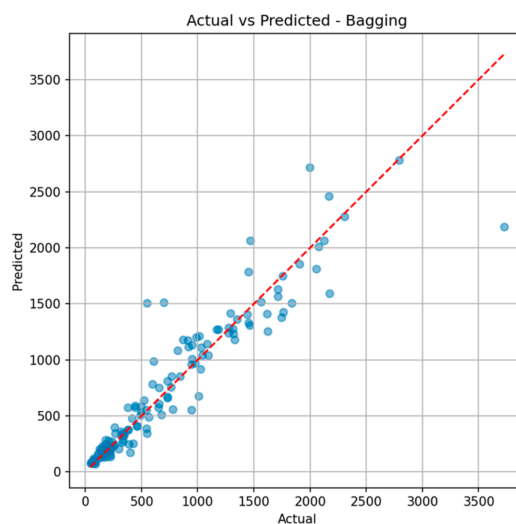
(c)



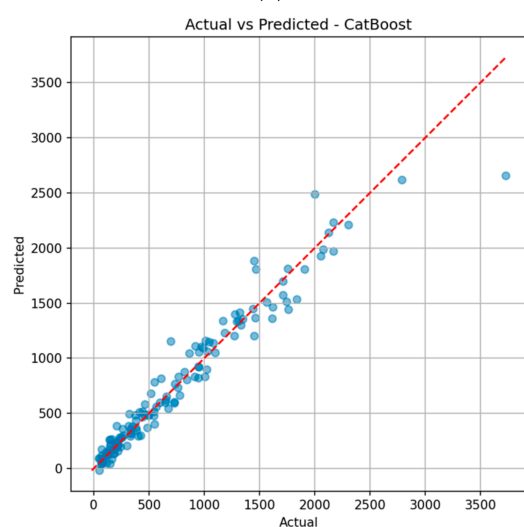
(d)



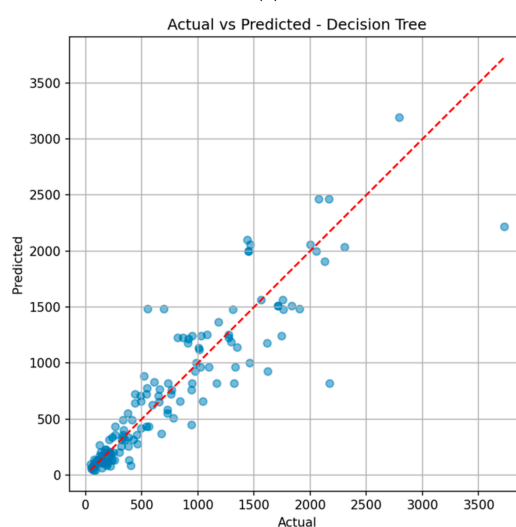
(e)



(f)



(g)



(h)

Figure 3. Cont.

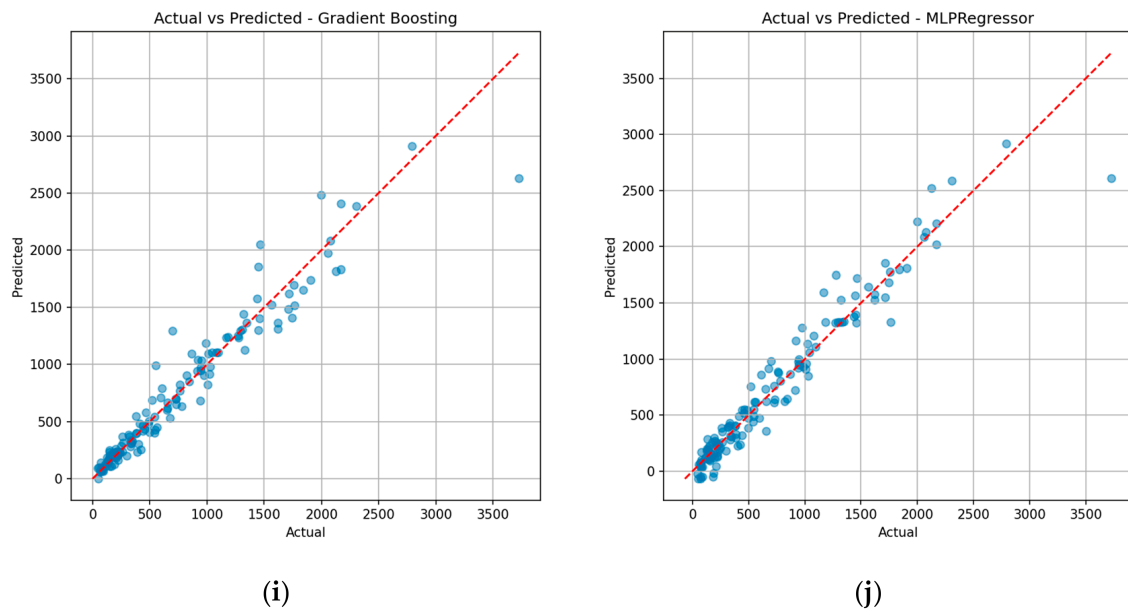


Figure 3. Scatter plots illustrating the correlation between actual and predicted values for the analyzed machine learning algorithms (a–j). The dashed line represents the perfect fit (1:1). The individual plots correspond to the models: (a) Random Forest, (b) Stacking Regressor, (c) Voting Regressor, (d) XGBoost, (e) AdaBoost, (f) Bagging, (g) CatBoost, (h) Decision Tree, (i) Gradient Boosting, (j) MLPRegressor.

3.3. Overfitting Diagnostics

As part of the analysis, differences in the R^2 coefficient between the training, validation, and test datasets were examined. A reference value was established to reflect an ideal scenario without overfitting. In the context of the discussed problem, most models showed a decrease in the R^2 coefficient when transitioning from the training dataset to the validation/test dataset. This may indicate that the model attempted to fit the training data. An important aspect of this process was ensuring that the decline was not so drastic as to potentially lead to overfitting. On the chart indicating overfitting between the training and test sets, it can be observed that the decision tree exhibits the greatest difference compared to other models. The literature notes that decision trees can overfit much faster, but to mitigate this phenomenon, hyperparameters such as `max_depth` and `min_samples_leaf` were used [79]. The difference between validation and test datasets suggests that the decision tree model tends to overfit more than other models (Figure 4). The models were designed and trained using cross-validation and hyperparameter optimization aimed at reducing overfitting. Most of them show a reasonable difference between training and testing results, and validation results closely reflected test outcomes. It appears that only the mentioned decision tree exhibits the greatest overfitting. The overfitting chart helps to understand how reliable the gross CO₂ emission predictions are and whether the models have learned by memorization.

3.4. Learning Curve

The description discusses the learning curve, which tracks the model's performance as the number of training samples grows during training (Figure 5a–j). The X-axis represents the number of training data samples, while the Y-axis shows the average RMSE (root mean square error), determined through 5-fold cross-validation. The number of samples in this research problem describes the size of the training (learning) set. In the study, it directly corresponds to the records included in the cleaned, secondary data set (`df_cleaned`). A slight increase or stabilization of the red training error curve suggested that the models

are no longer merely memorizing the data but are beginning to focus on patterns, which is important for reducing overfitting. A systematic decrease in the green validation curve indicated that, with the influx of new information from the dataset, the algorithm was increasingly better at predicting unseen data. The convergence of both lines with minimal difference between them confirms the achievement of high generalization ability and stability of the developed machine learning algorithm. The CatBoost and XGBoost models achieved ideal stabilization of the training error (Figure 5d,g). The other learning algorithms also aimed to stabilize the training error. Additionally, each of the planned models aimed to reduce the validation curve. The results also confirm, based on the previously mentioned quality metrics of the model, that Decision Tree had the largest difference (Figure 5h), while MLPRegressor had the smallest difference (Figure 5j), with both lines approaching each other.

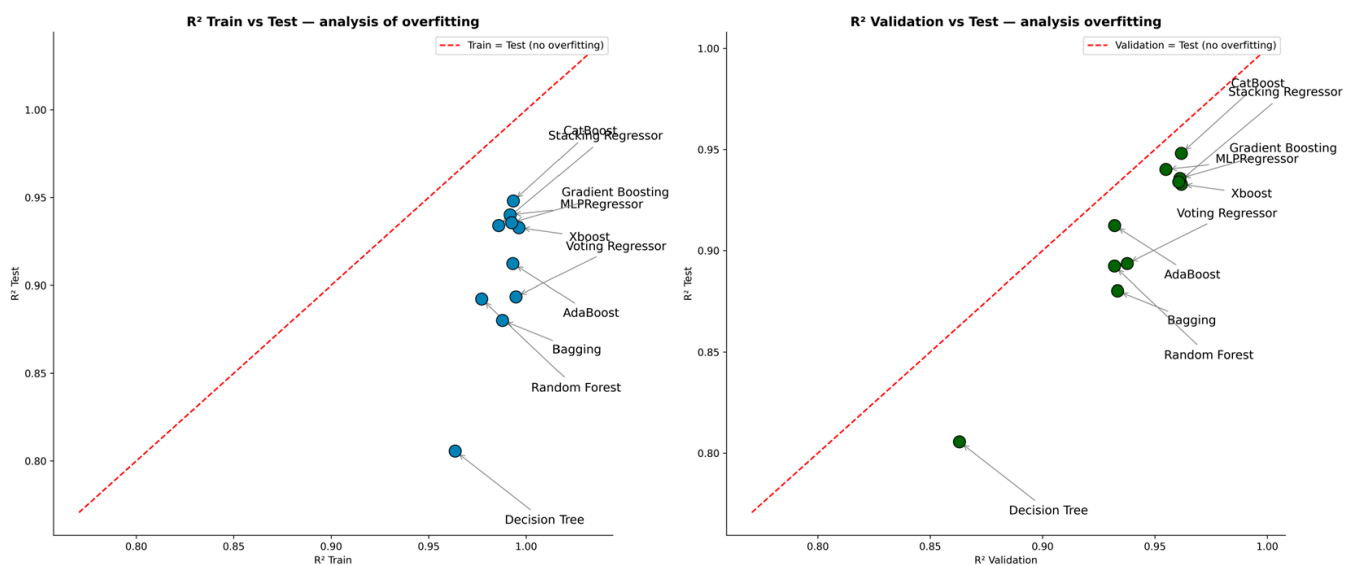


Figure 4. Assessment of the generalization ability and risk analysis of overfitting in the studied regression models. Dashed lines (Train = Test and Validation = Test) serve as reference axes, representing the theoretical state of perfect generalization. The proximity of data points to these lines indicates minimal overfitting and high stability of predictions on unseen data.

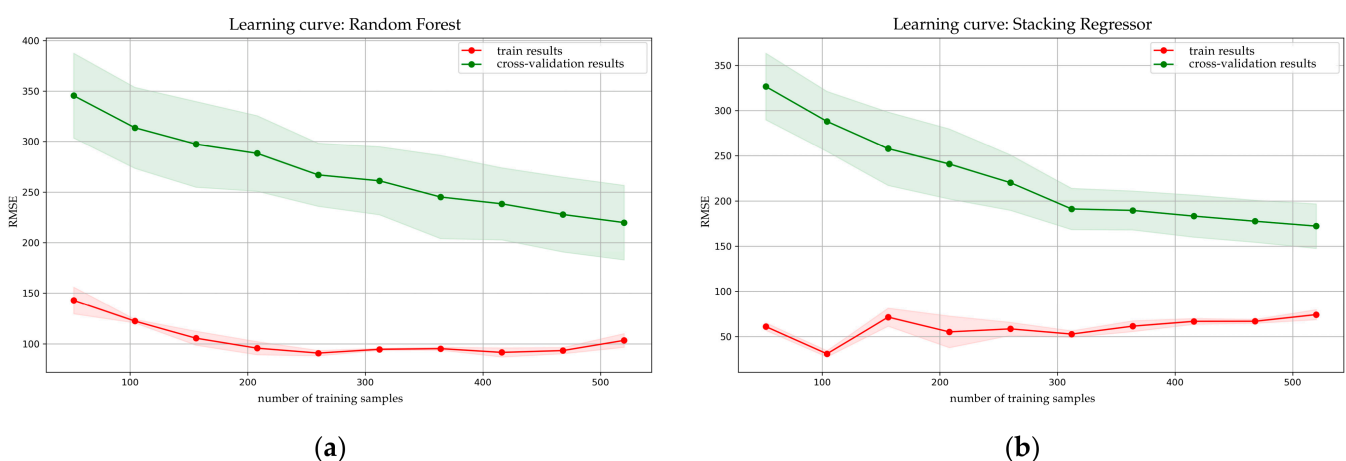


Figure 5. Cont.

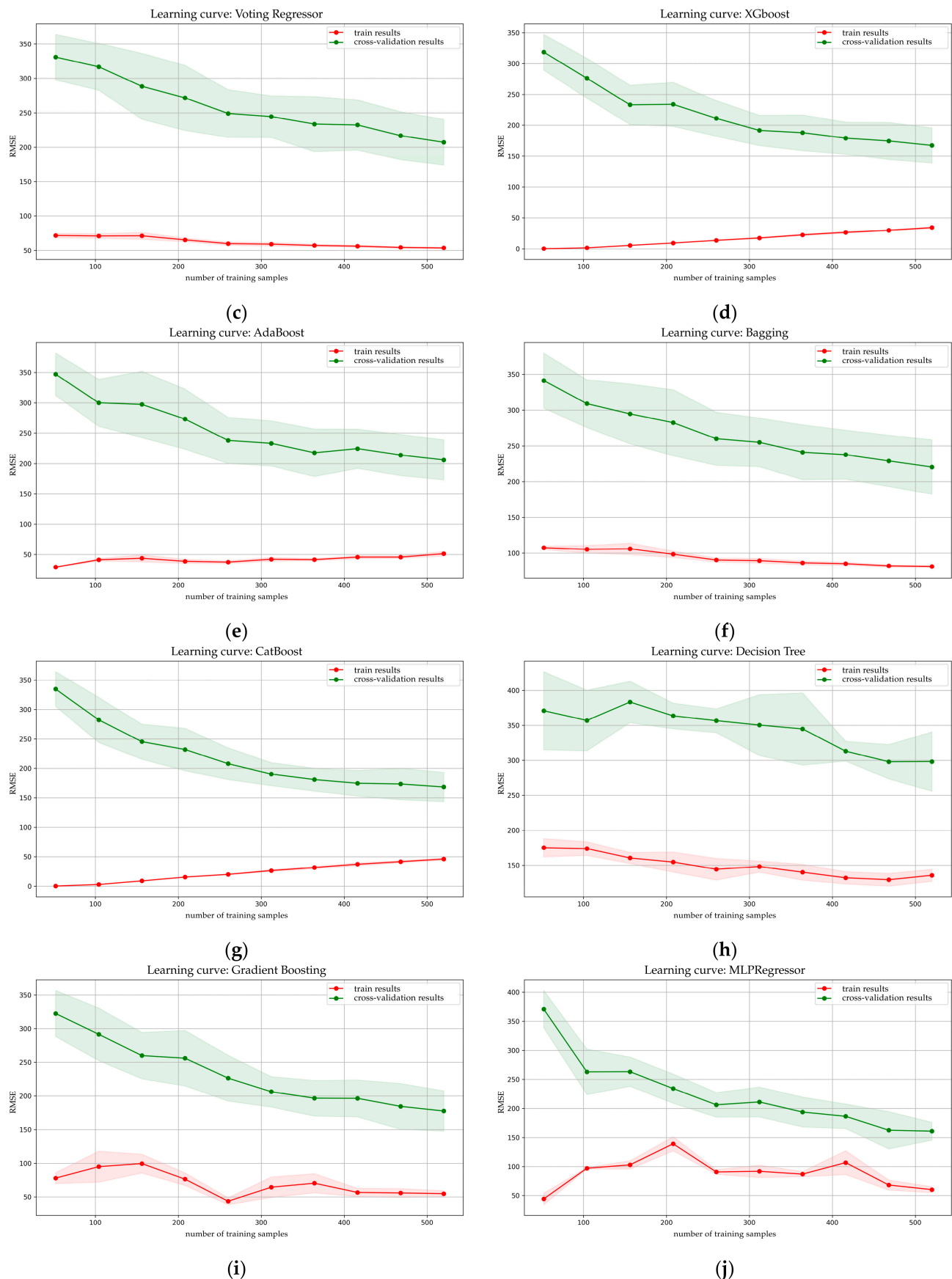


Figure 5. Learning curves for analyzed machine learning algorithms. The individual plots correspond to the models: (a) Random Forest, (b) Stacking Regressor, (c) Voting Regressor, (d) XGBoost, (e) AdaBoost, (f) Bagging, (g) CatBoost, (h) Decision Tree, (i) Gradient Boosting, (j) MLPRegressor.

3.5. Explainability of Variables by SHAP

The final stage of the analysis was dedicated to interpreting the results using explainable artificial intelligence (XAI) methods, employing the SHAP (SHapley Additive exPlanations) algorithm. This method allowed for the decomposition of predictions and precise assessment of the marginal contribution of each independent variable to the final model output [67–69,80,81].

The analysis conducted for hybrid models enabled a detailed interpretation of the mechanisms controlling the forecasts of gross CO₂ balance in relation to environmental parameters (Figure 6a–t). On one hand, the bar chart (representing the global interpretation) clearly verifies the absolute hierarchy of the top 10 most important variables, placing height, age, and number of trees at the top, which serves as the main starting point for understanding the algorithm’s priorities. On the other hand, the scatter plot (point interpretation) accurately explains the mechanics of these features, demonstrating how dynamically and in what manner they influence specific predictions. It was shown that the variable “Height-estimated data [m]” is the most significant factor affecting the carbon balance, which is reflected in the literature, where SHAP regularly identifies important physical parameters (such as wave height *H_s* in hydrological studies or rainfall in agriculture) as the main drivers of the models. In the case of the MLPRegressor model, the dominant role was “Age (years)” (Figure 6j). The best Catboost model based its predictions mainly on the first variable, “Height-estimated data [m]”, whose SHAP values range asymmetrically from −500 to +1000. The second variable, “Age (years)”, shows a symmetric influence, ranging from −250 to approximately +700, suggesting that high values of this feature significantly increase the model’s output more than low values decrease it. The third variable, „Number of trees-estimated data [ha]”, although ranked third in importance, is particularly interesting. Its negative impact reaches −750, indicating that low values of this feature can most substantially decrease the result. The colour scheme of the chart, from blue to red, confirmed a clear directional relationship: low values push the prediction down, high values push it up. The remaining variables in CatBoost have SHAP values ranging from −100 to +100, indicating their marginal contribution and suggesting they can be treated as weak signals or noise. In summary, the three aforementioned variables in CatBoost determine most of the model’s predictive power, with the dominant first and third variables requiring attention due to their potential negative impact. Models such as CatBoost, XGBoost, Gradient Boosting, and others focused on estimating tree height, which was one of the three most important factors in their analyses. Tree age was also a significant variable across all the rankings, as confirmed by the results, particularly for MLPRegressor. These results demonstrate high physical and substantive consistency.

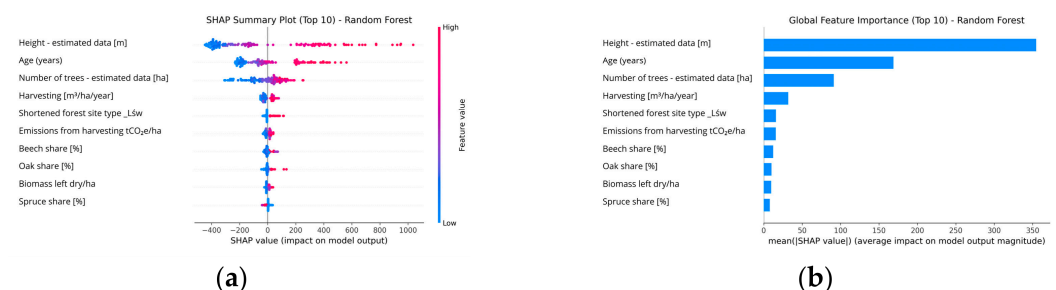


Figure 6. Cont.

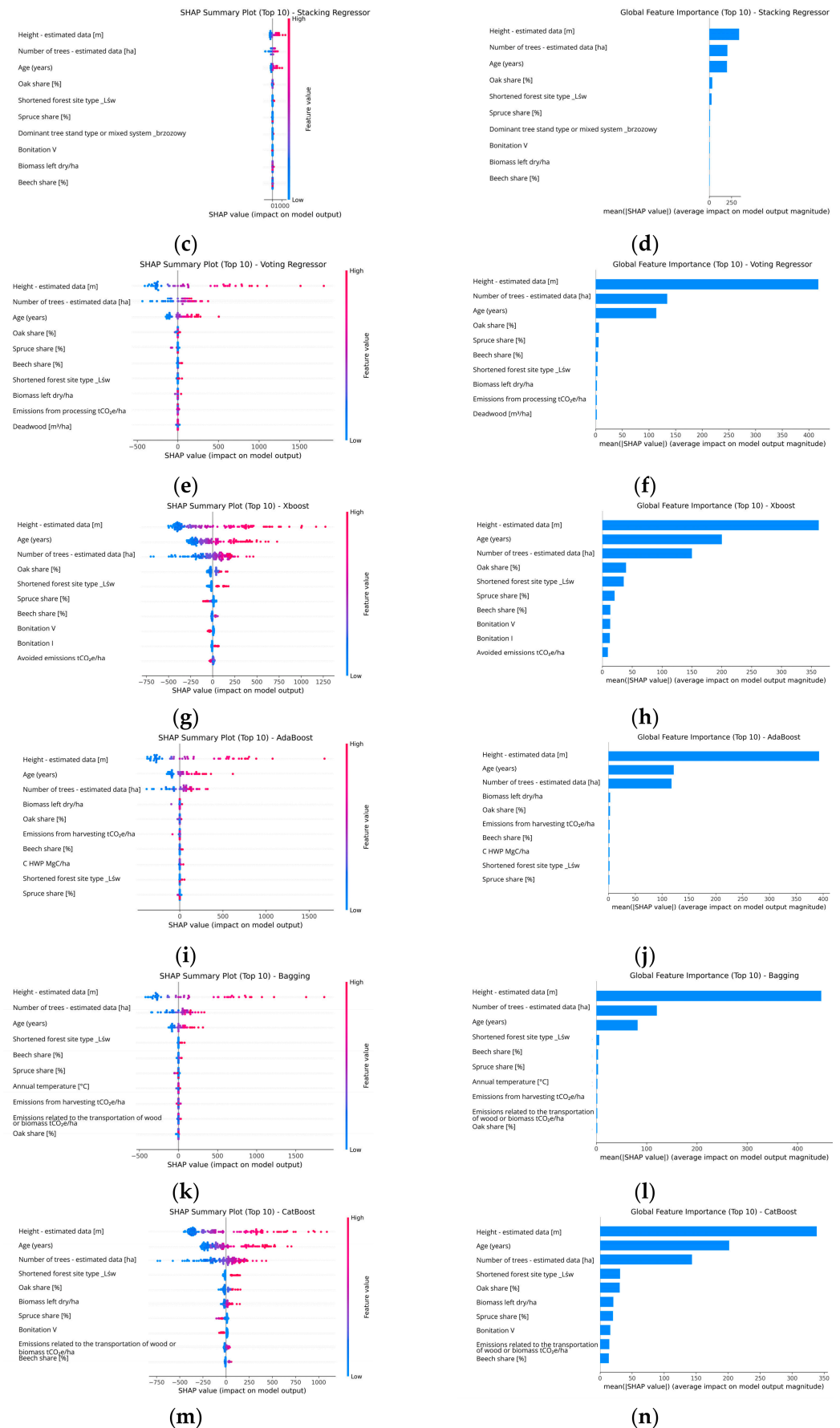


Figure 6. Cont.

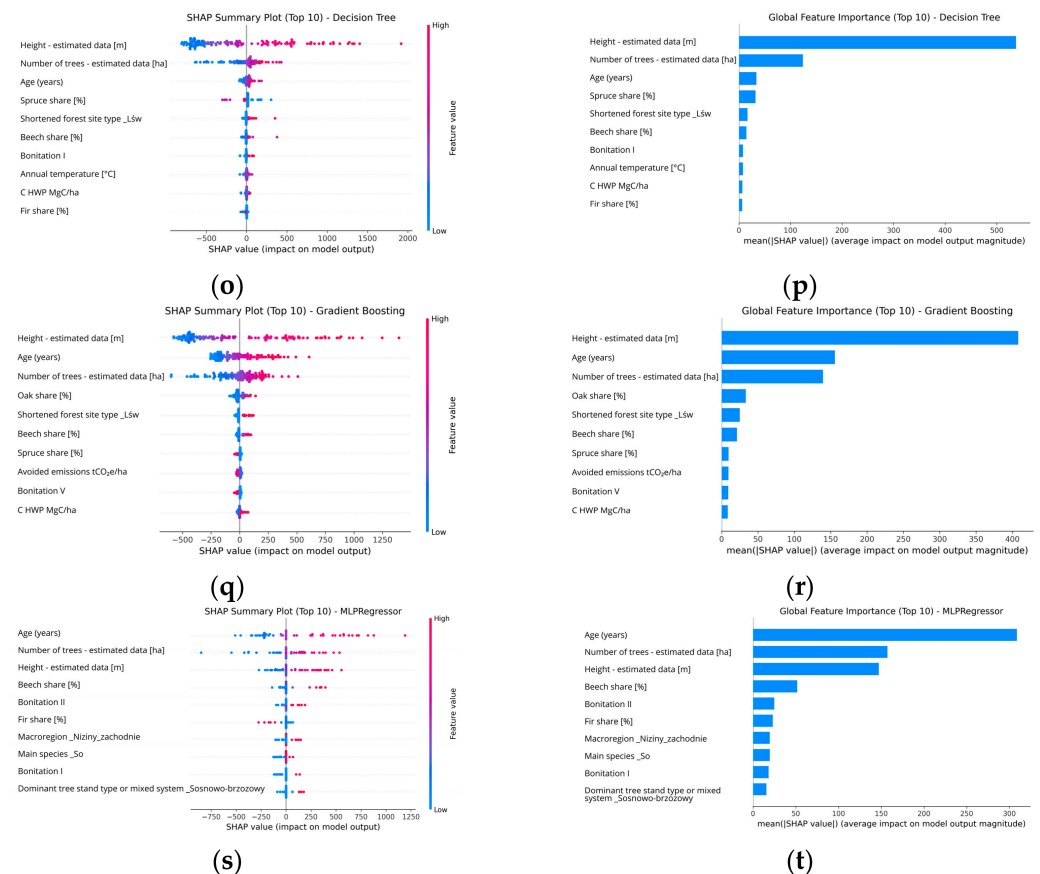


Figure 6. SHAP for the analyzed machine learning algorithms: (a,b) Random Forest, (c,d) Stacking Regressor, (e,f) Voting Regressor, (g,h) XGBoost, (i,j) AdaBoost, (k,l) Bagging, (m,n) CatBoost, (o,p) Decision Tree, (q,r) Gradient Boosting, (s,t) MLPRegressor.

Integrating SHAP analysis with black-box models allowed for the identification of the most effective set of variables while maintaining record-breaking algorithm accuracy. The literature emphasizes that this approach not only simplifies the model structure but also is crucial for building trust in the decisions made by artificial intelligence [79].

The results obtained in this study indicate that the gross CO₂e balance in the analyzed forest stands can be predicted with very high accuracy using machine learning models based on habitat, stand, and operational data (see Tables 1 and 3; Figure 2). It is particularly significant that the highest effectiveness was achieved by models capable of capturing complex nonlinear relationships, primarily CatBoost, MLPRegressor, Gradient Boosting, XGBoost, and Stacking Regressor. This confirms that, in the examined problem, the relationship structure between input features and gross CO₂e is multivariate and not well described by simpler single-tree models [11,12]. An equally important result was the consistency between high prediction accuracy and the models' generalization ability (Figures 3 and 4). Analysis of differences between training, validation, and test results, as well as the learning curve progression, indicates that most models maintained stability on unseen data [18]. The weakest results obtained for Decision Tree suggest a greater susceptibility of this algorithm to overfitting and data structure simplification. From a practical perspective, this means that in applications related to carbon balance assessment, hybrid and nonlinear models are more useful than algorithms with simpler architectures [12].

Furthermore, the interpretation of SHAP showed that the most important predictor of gross CO₂e was tree height (Height-estimated data [m]), followed by tree age and number of trees (Figure 6). This arrangement of results is substantively consistent with the ecological logic of forest stand functioning, as these features directly reflect biomass

accumulation potential, resource structure, and growth conditions [3,81]. The obtained results thus indicate that modelling gross CO₂e should not be limited to single carbon pools but should also consider structural, habitat, and economic parameters [5,7]. In this sense, the presented approach has value not only for prediction but also for interpretation and application within forest management and carbon inventory contexts.

From a practical standpoint, the results suggest that the development of predictive models for gross CO₂e should integrate multi-source data, including stand, habitat, environmental, operational, and product parameters. Approaches combining inventory data with algorithms capable of modelling nonlinear relationships, while maintaining good generalization, scalability, and interpretability, are most valuable [18]. Such models could serve as the basis for tools supporting scenario assessments, carbon balance estimation, and forest management planning under circular-economy conditions [8,10]. Their utility could be further enhanced by incorporating remote sensing data and information on environmental variability over time [23]. Simultaneously, the results indicate that further development of these tools should include external validation, greater regional differentiation, and the integration of remote sensing and dynamic environmental indicators. Such an approach could increase model robustness to changing conditions and broaden its applicability in scenario analyses, climate reporting, and evaluating the forest-wood sector's contribution to achieving the aim of climate neutrality.

4. Conclusions

This research involved developing and optimizing machine learning models to predict the gross CO₂e balance in Polish forest stands using habitat, stand, carbon, and operational data. The approach combined an assessment of predictive performance with an analysis of generalization ability and an interpretation of variable impacts using SHAP. The results confirmed the high predictive effectiveness of nonlinear and hybrid models, with the best results achieved by CatBoost, MLPRegressor, Gradient Boosting, XGBoost and Stacking Regressor. Among ten tested algorithms, the highest predictive performance was obtained by the CatBoost model, with an R² of 0.9481 and the lowest prediction error expressed by RMSE. Six models achieved an R² above 0.90, confirming their usefulness in modelling complex environmental dependencies. The analysis also showed that key factors influencing the gross CO₂e were tree height, tree age, number of trees, species composition, and selected habitat features. The main conclusions are as follows:

- The gross CO₂e balance can be predicted with high accuracy based on variables available in forestry practice;
- The most useful models were those capable of capturing non-linear relationships and interactions between variables;
- Simpler single-tree models proved more susceptible to overfitting and less stable on unseen data;
- SHAP interpretation confirmed the dominant role of structural features in shaping the gross CO₂e;
- Reliable assessment of gross CO₂e requires simultaneous consideration of biological, habitat, operational, and product components;
- Further development of models should include external validation, greater regional differentiation, and integration of remote sensing data and dynamic environmental indicators.

The proposed approach has both cognitive and practical significance. It demonstrates that predictive models can support accurate, interpretable, and decision-useful assessments of forest carbon balances. In this sense, it provides a solid foundation for developing more advanced tools for forestry and circular economy management. Future research will focus on incorporating hyperspectral and satellite imaging data to monitor biomass and forest

health. In the era of digital transformation, unmanned aerial vehicles equipped with sensors could provide data on canopy structure and vegetation indices, enhancing the predictive capacity of above-ground biomass models. Additionally, further use of explainable AI methods is planned to ensure transparency in AI decision-making processes.

Author Contributions: Conceptualization, K.P. (Krzysztof Przybył); data curation, K.P. (Krzysztof Pilarski); formal analysis, K.P. (Krzysztof Przybył); investigation, K.P. (Krzysztof Przybył) and K.P. (Krzysztof Pilarski); methodology, K.P. (Krzysztof Przybył) and K.P. (Krzysztof Pilarski); project administration, K.P. (Krzysztof Przybył) and A.A.P.; resources, A.A.P. and K.P. (Krzysztof Pilarski); software, K.P. (Krzysztof Przybył); supervision, K.P. (Krzysztof Przybył); validation, K.P. (Krzysztof Przybył); visualization, K.P. (Krzysztof Przybył); writing—original draft preparation, K.P. (Krzysztof Przybył) and A.A.P.; writing—review and editing, K.P. (Krzysztof Przybył), A.A.P. and K.P. (Krzysztof Pilarski). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data supporting the findings of this study are contained within the article.

Acknowledgments: During the preparation of this manuscript, the authors used Canva (canva.com) to assist by design and prepare graphical elements, and Grammarly (grammarly.com) to assist with translation and language correction. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cienciala, E.; Melichar, J. Forest Carbon Stock Development Following Extreme Drought-Induced Dieback of Coniferous Stands in Central Europe: A CBM-CFS3 Model Application. *Carbon. Balance Manag.* **2024**, *19*, 1. [\[CrossRef\]](#)
2. Saponaro, V.; De Cáceres, M.; Dalmonech, D.; D’Andrea, E.; Vangi, E.; Collalti, A. Assessing the Combined Effects of Forest Management and Climate Change on Carbon and Water Fluxes in European Beech Forests. *For. Ecosyst.* **2025**, *12*, 100290. [\[CrossRef\]](#)
3. Pretzsch, H.; Hilmers, T. Structural Diversity and Carbon Stock of Forest Stands: Tradeoff as Modified by Silvicultural Thinning. *Eur. J. For. Res.* **2025**, *144*, 775–796. [\[CrossRef\]](#)
4. Araza, A.; de Bruin, S.; Herold, M.; Quegan, S.; Labriere, N.; Rodriguez-Veiga, P.; Avitabile, V.; Santoro, M.; Mitchard, E.T.A.; Ryan, C.M.; et al. A Comprehensive Framework for Assessing the Accuracy and Uncertainty of Global Above-Ground Biomass Maps. *Remote Sens. Environ.* **2022**, *272*, 112917. [\[CrossRef\]](#)
5. Wei, X.; Zhao, J.; Hayes, D.J.; Daigneault, A.; Zhu, H. A Life Cycle and Product Type Based Estimator for Quantifying the Carbon Stored in Wood Products. *Carbon. Balance Manag.* **2023**, *18*, 1. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Bianchi, M.; Cascavilla, A.; Diaz, J.C.; Ladu, L.; Blazquez, B.P.; Pierre, M.; Staffieri, E.; Yilan, G. Circular Bioeconomy: A Review of Empirical Practices across Implementation Scales. *J. Clean. Prod.* **2024**, *477*, 143816. [\[CrossRef\]](#)
7. Xie, S.H.; Kurz, W.A.; Smyth, C.; Xu, Z.; Roeser, D. Forest Products Circular Economy in an Export-Focused Jurisdiction: Can It Fill the Emission Reduction Gap? *Clean. Circ. Bioeconomy* **2024**, *8*, 100096. [\[CrossRef\]](#)
8. Szichta, P.; Risse, M.; Weber-Blaschke, G.; Richter, K. Environmental Potentials from Wood Cascading: A Future-Oriented Consequential yet Dynamic Approach Considering Market and Time-Dependent Biogenic Carbon Effects for Selected Scenarios under German Conditions. *Clean. Circ. Bioeconomy* **2024**, *9*, 100103. [\[CrossRef\]](#)
9. Wedajo, D.Y.; Cristescu, C.; Billore, S.; Adamopoulos, S. Carbon Impact of Wood-Based Products through Substitution: A Review of Assessment Aspects and Future Research Perspectives in Life Cycle Assessment. *Carbon Manag.* **2025**, *16*, 2536350. [\[CrossRef\]](#)
10. Hoppen, M.; Baier, S.; Schinke, L.; Ziesak, M.; Schreiber, L.J.; Wahl, A.; Chen, J.; Bektas, A.R.; Heinze, F.; Schluse, M.; et al. Digital Technologies for Precise Carbon Balancing in Timber Procurement. *Eur. J. For. Res.* **2025**, *144*, 1043–1061. [\[CrossRef\]](#)
11. Zhao, C.; Zhang, M.; Bai, J.; Wu, J.; Chang, I.-S. A Review of the Application of Machine Learning in Carbon Emission Assessment Studies: Prediction Optimization and Driving Factor Selection. *Sci. Total Environ.* **2025**, *987*, 179678. [\[CrossRef\]](#)
12. Fasihi, M.; Portelli, B.; Cadez, L.; Tomao, A.; Falcon, A.; Alberti, G.; Serra, G. Assessing Ensemble Models for Carbon Sequestration and Storage Estimation in Forests Using Remote Sensing Data. *Ecol. Inform.* **2024**, *83*, 102828. [\[CrossRef\]](#)

13. Dadgar, M.; Faramarzi, S.E. Assessing the Performance of Machine Learning Models for Predicting Soil Organic Carbon Variability across Diverse Landforms. *Environ. Earth Sci.* **2024**, *83*, 657. [CrossRef]
14. Li, T.; Cui, L.; Kuhnert, M.; McLaren, T.I.; Pandey, R.; Liu, H.; Wang, W.; Xu, Z.; Xia, A.; Dalal, R.C.; et al. A Comprehensive Review of Soil Organic Carbon Estimates: Integrating Remote Sensing and Machine Learning Technologies. *J. Soils Sediments* **2024**, *24*, 3556–3571. [CrossRef]
15. Li, Y.; Li, J.; Tan, J.; Ma, T.; Yan, X.; Chen, Z.; Li, K. Fine Resolution Mapping of Forest Soil Organic Carbon Based on Feature Selection and Machine Learning Algorithm. *Remote Sens.* **2025**, *17*, 2000. [CrossRef]
16. Triantakoustantis, D.; Karakostas, A. Soil Organic Carbon Monitoring and Modelling via Machine Learning Methods Using Soil and Remote Sensing Data. *Agriculture* **2025**, *15*, 910. [CrossRef]
17. Padalia, H.; Prakash, A.; Watham, T. Modelling Aboveground Biomass of a Multistage Managed Forest through Synergistic Use of Landsat-OLI, ALOS-2 L-Band SAR and GEDI Metrics. *Ecol. Inform.* **2023**, *77*, 102234. [CrossRef]
18. Zheng, M.; Wen, Q.; Xu, F.; Wu, D. Regional Forest Carbon Stock Estimation Based on Multi-Source Data and Machine Learning Algorithms. *Forests* **2025**, *16*, 420. [CrossRef]
19. Lech, P.; Hildebrand, R.; Małachowska, J. Forest Monitoring in Poland: Legal Foundations and Scope of the Programme. *Folia For. Pol.* **2025**, *67*, 35–45. [CrossRef]
20. National Forest Inventory-BULiGL EN. Available online: <https://buligl.pl/pl/web/buligl-en/w/national-forest-inventory> (accessed on 10 April 2026).
21. Integrating Digital Technologies for Comprehensive Carbon Accounting in Forests and Agroforestry Systems. Available online: <https://openpub.fmach.it/handle/10449/88285> (accessed on 14 May 2026).
22. Ali, A.; Russell, J.D. Accelerating the Transition to Wood-Based Circular Bioeconomy: A Literature Review of Current State, Trends, Opportunities, and Priorities for Future Research. *Curr. For. Rep.* **2025**, *11*, 23. [CrossRef]
23. Grabska-Szwagrzyk, E.; Tiede, D.; Sudmanns, M.; Kozak, J. Map of Forest Tree Species for Poland Based on Sentinel-2 Data. *Earth Syst. Sci. Data* **2024**, *16*, 2877–2891. [CrossRef]
24. Assefa, G.; Mengistu, T.; Getu, Z.; Zewdie, S. *Training Manual on Forest Carbon Pools and Carbon Stock Assessment in the Context of Sustainable Forest Management and REDD+*; Wondo Genet; Hawassa University: Hawassa, Ethiopia, 2013.
25. Garba, M.; Usman, M.; Saidu, M. Enhancing employee attrition prediction: The impact of data preprocessing on machine learning model performance. *FUDMA J. Sci.* **2025**, *9*, 205–210. [CrossRef]
26. Song, F.; Liu, H.; Ma, H.; Chen, X.; Wang, S.; Qin, T.; Liang, H.; Huang, D. AI Model Based on Diaphragm Ultrasound to Improve the Predictive Performance of Invasive Mechanical Ventilation Weaning: Prospective Cohort Study. *JMIR Form. Res.* **2025**, *9*, e72482. [CrossRef] [PubMed]
27. Guo, S.; Wang, Z.; Liang, S. Calculation and Analysis of Load DC Magnetic Bias of Three-Phase Five-Column Transformer. *High Volt. Appar.* **2023**, *59*, 113–121+129. [CrossRef]
28. Prediksi Dan Deteksi Bug Pada Software Menggunakan Pendekatan Machine Learning. Available online: <https://bpika.uma.ac.id/2025/12/02/penerapan-machine-learning-dalam-software-engineering-untuk-prediksi-bug/> (accessed on 14 May 2026).
29. Przybył, K.; Koszela, K. Applications MLP and Other Methods in Artificial Intelligence of Fruit and Vegetable in Convective and Spray Drying. *Appl. Sci.* **2023**, *13*, 2965. [CrossRef]
30. Przybył, K.; Masewicz, Ł.; Koszela, K.; Duda, A.; Szychta, M.; Gierz, Ł. An MLP Artificial Neural Network for Detection of the Degree of Saccharification of Arabic Gum Used as a Carrier Agent of Raspberry Powders. In *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*; SPIE: Bellingham, WA, USA, 2021; Volume 11878, pp. 605–609. [CrossRef]
31. Orrù, P.F.; Zoccheddu, A.; Sassu, L.; Mattia, C.; Cozza, R.; Arena, S. Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry. *Sustainability* **2020**, *12*, 4776. [CrossRef]
32. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. *Neurocomputing* **2022**, *503*, 92–108. [CrossRef]
33. Ogundokun, R.O.; Maskeliunas, R.; Misra, S.; Damaševičius, R. Improved CNN Based on Batch Normalization and Adam Optimizer. In *Computational Science and Its Applications*; Springer International Publishing: Cham, Switzerland, 2022; LNCS; Volume 13381, pp. 593–604. [CrossRef]
34. Yang, J.; Yang, G. Modified Convolutional Neural Network Based on Dropout and the Stochastic Gradient Descent Optimizer. *Algorithms* **2018**, *11*, 28. [CrossRef]
35. Yulisa, A.; Park, S.H.; Choi, S.; Chairattanawat, C.; Hwang, S. Enhancement of Voting Regressor Algorithm on Predicting Total Ammonia Nitrogen Concentration in Fish Waste Anaerobiosis. *Waste Biomass Valorization* **2023**, *14*, 461–478. [CrossRef]
36. Chen, S.; Zheng, W. RRMSE-Enhanced Weighted Voting Regressor for Improved Ensemble Regression. *PLoS ONE* **2025**, *20*, e0319515. [CrossRef]
37. Ahmad, A.; Farooq, F.; Niewiadomski, P.; Ostrowski, K.; Akbar, A.; Aslam, F.; Alyousef, R. Prediction of Compressive Strength of Fly Ash Based Concrete Using Individual and Ensemble Algorithm. *Materials* **2021**, *14*, 794. [CrossRef] [PubMed]

38. Nacar, E.N.; Erdebilli, B.; Eraslan, E. Toward Green Manufacturing: A Heuristic Hybrid Machine Learning Framework with PSO for Scrap Reduction. *Sustainability* **2025**, *17*, 9106. [\[CrossRef\]](#)
39. Erbulut, Ö.G.; Çolak, Z. A Hybrid Machine Learning Approach for Housing Price Prediction: The Stacking Regressor Method. *Int. J. Hous. Mark. Anal.* **2026**, *19*, 942–970. [\[CrossRef\]](#)
40. Madhukesh, J.K.; Madhu, J.; Fareeduddin, M.; Chandan, K.; Khan, U.; Al-Tref, G.A.; Hussain, S.M.; Nagaraja, K.V.; Kumar, R. Implementation of Stacking Regressor Model on the Flow Induced by TiO₂-H₂O and Ti₆Al₄V-H₂O Nanofluid with Waste Discharge Concentration. *ZAMM Z. Fur Angew. Math. Und Mech.* **2024**, *104*, e202300796. [\[CrossRef\]](#)
41. Aslam, F.; Alyousef, R.; Hassan Awan, H.; Faisal Javed, M. Forecasting the Self-Healing Capacity of Engineered Cementitious Composites Using Bagging Regressor and Stacking Regressor. *Structures* **2023**, *54*, 1717–1728. [\[CrossRef\]](#)
42. Mahamat, A.A.; Boukar, M.M.; Leklou, N.; Celino, A.; Obianyo, I.I.; Bih, N.L.; Stanislas, T.T.; Savastanos, H. Decision Tree Regression vs. Gradient Boosting Regressor Models for the Prediction of Hygroscopic Properties of Borassus Fruit Fiber. *Appl. Sci.* **2024**, *14*, 7540. [\[CrossRef\]](#)
43. Bagalkot, N.; Keprate, A.; Orderløkken, R. Combining Computational Fluid Dynamics and Gradient Boosting Regressor for Predicting Force Distribution on Horizontal Axis Wind Turbine. *Vibration* **2021**, *4*, 17. [\[CrossRef\]](#)
44. Li, X.; Li, W.; Xu, Y. Human Age Prediction Based on DNA Methylation Using a Gradient Boosting Regressor. *Genes* **2018**, *9*, 424. [\[CrossRef\]](#)
45. Sharma, H.; Harsora, H.; Ogunleye, B. An Optimal House Price Prediction Algorithm: XGBoost. *Analytics* **2024**, *3*, 30–45. [\[CrossRef\]](#)
46. Niazkar, M.; Menapace, A.; Brentan, B.; Piraei, R.; Jimenez, D.; Dhawan, P.; Righetti, M. Applications of XGBoost in Water Resources Engineering: A Systematic Literature Review (Dec 2018–May 2023). *Environ. Model. Softw.* **2024**, *174*, 105971. [\[CrossRef\]](#)
47. Zhang, P.; Jia, Y.; Shang, Y. Research and Application of XGBoost in Imbalanced Data. *Int. J. Distrib. Sens. Netw.* **2022**, *18*, 15501329221106935. [\[CrossRef\]](#)
48. Hakkal, S.; Lahcen, A.A. XGBoost To Enhance Learner Performance Prediction. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100254. [\[CrossRef\]](#)
49. Oukhouya, H.; Kadiri, H.; El Himdi, K.; Guerbaz, R. Forecasting International Stock Market Trends: XGBoost, LSTM, LSTM-XGBoost, And Backtesting XGBoost Models. *Stat. Optim. Inf. Comput.* **2024**, *12*, 200–209. [\[CrossRef\]](#)
50. Guan, X.; Xue, R.; He, Z.; Chen, S.; Chen, X. CatBoost-Optimized Hyperspectral Modeling for Accurate Prediction of Wood Dyeing Formulations. *Forests* **2025**, *16*, 1279. [\[CrossRef\]](#)
51. Elmasry, N.H.; Elshaarawy, M.K. Hybrid Metaheuristic Optimized Catboost Models for Construction Cost Estimation of Concrete Solid Slabs. *Sci. Rep.* **2025**, *15*, 21612. [\[CrossRef\]](#)
52. Hadianito, A.; Utomo, W.H. CatBoost Optimization Using Recursive Feature Elimination. *J. Online Inform.* **2024**, *9*, 169–178. [\[CrossRef\]](#)
53. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for Big Data: An Interdisciplinary Review. *J. Big Data* **2020**, *7*, 94. [\[CrossRef\]](#)
54. Przybył, K. Explainable AI: Machine Learning Interpretation in Blackcurrant Powders. *Sensors* **2024**, *24*, 3198. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Alghamdi, S.J. Classifying High Strength Concrete Mix Design Methods Using Decision Trees. *Materials* **2022**, *15*, 1950. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Przybył, K.; Walkowiak, K.; Kowalczewski, P.L. Efficiency of Identification of Blackcurrant Powders Using Classifier Ensembles. *Foods* **2024**, *13*, 697. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Hamzat, A.K.; Salman, U.T.; Murad, M.S.; Altay, O.; Bahceci, E.; Asmatulu, E.; Bakir, M.; Asmatulu, R. Predicting Flexural Strengths of Fiber-Reinforced Polymeric Composites. *Hybrid. Adv.* **2025**, *8*, 100385. [\[CrossRef\]](#)
58. Mosavi, A.; Sajedi Hosseini, F.; Choubin, B.; Goodarzi, M.; Dineva, A.A.; Rafiei Sardooi, E. Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction. *Water Resour. Manag.* **2021**, *35*, 23–37. [\[CrossRef\]](#)
59. Yılmaz, Y.; Nayır, S. Machine Learning Based Prediction of Compressive and Flexural Strength of Recycled Plastic Waste Aggregate Concrete. *Structures* **2024**, *69*, 107363. [\[CrossRef\]](#)
60. Fuchs, M.; Krautenbacher, N. Minimization and Estimation of the Variance of Prediction Errors for Cross-Validation Designs. *J. Stat. Theory Pract.* **2016**, *10*, 420–443. [\[CrossRef\]](#)
61. Bengio, Y.; Grandvalet, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J. Mach. Learn. Res.* **2004**, *5*, 1089–1105.
62. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [\[CrossRef\]](#)
63. Karch, J. Improving on Adjusted R-Squared. *Collabra Psychol.* **2020**, *6*, 45. [\[CrossRef\]](#)
64. Weng, S.; Yu, S.; Guo, B.; Tang, P.; Liang, D. Non-Destructive Detection of Strawberry Quality Using Multi-Features of Hyperspectral Imaging and Multivariate Methods. *Sensors* **2020**, *20*, 3074. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Wang, Y.; Zhang, W.; Gao, R.; Jin, Z.; Wang, X. Recent Advances in the Application of Deep Learning Methods to Forestry. *Wood Sci. Technol.* **2021**, *55*, 1171–1202. [\[CrossRef\]](#)
66. Nohara, Y.; Matsumoto, K.; Soejima, H.; Nakashima, N. Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital. *Comput. Methods Programs Biomed.* **2022**, *214*, 106584. [\[CrossRef\]](#) [\[PubMed\]](#)

67. Wadoux, A.M.J.C.; Heuvelink, G.B.M.; de Bruin, S.; Brus, D.J. Spatial Cross-Validation Is Not the Right Way to Evaluate Map Accuracy. *Ecol. Modell.* **2021**, *457*, 109692. [[CrossRef](#)]
68. Han, J.; Guzman, J.A.; Chu, M.L. Prediction of Gully Erosion Susceptibility through the Lens of the SHapley Additive ExPlanations (SHAP) Method Using a Stacking Ensemble Model. *J. Environ. Manag.* **2025**, *383*, 125478. [[CrossRef](#)]
69. Vega García, M.; Aznarte, J.L. Shapley Additive Explanations for NO₂ Forecasting. *Ecol. Inform.* **2020**, *56*, 101039. [[CrossRef](#)]
70. M'hamdi, O.; Takács, S.; Palotás, G.; Ilahy, R.; Helyes, L.; Pék, Z. A Comparative Analysis of XGBoost and Neural Network Models for Predicting Some Tomato Fruit Quality Traits from Environmental and Meteorological Data. *Plants* **2024**, *13*, 746. [[CrossRef](#)]
71. Martinović, M.; Dokic, K.; Pudić, D. Comparative Analysis of Machine Learning Models for Predicting Innovation Outcomes: An Applied AI Approach. *Appl. Sci.* **2025**, *15*, 3636. [[CrossRef](#)]
72. Przybył, K.; Gawalek, J.; Koszela, K.; Przybył, J.; Rudzińska, M.; Gierz, Ł.; Domian, E. Neural Image Analysis and Electron Microscopy to Detect and Describe Selected Quality Factors of Fruit and Vegetable Spray-Dried Powders—Case Study: Chokeberry Powder. *Sensors* **2019**, *19*, 4413. [[CrossRef](#)]
73. Yersaw, B.T.; Ebstu, E.T.; Areru, D.A.; Asres, L.A. Performance Evaluation of AquaCrop Model of Tomato under Stage Wise Deficit Drip Irrigation at Southern Ethiopia. *Adv. Agric.* **2024**, *2024*, 7201523. [[CrossRef](#)]
74. Hong, T.; Kim, C.J.; Jeong, J.; Kim, J.; Koo, C.; Jeong, K.; Lee, M. Framework for Approaching the Minimum CV(RMSE) Using Energy Simulation and Optimization Tool. *Proc. Energy Procedia* **2016**, *88*, 265–270. [[CrossRef](#)]
75. Wei, Y.; He, S.; Huang, P.; Duan, Y.; Dewanker, B.J.; Zhou, L. A Calibration Procedure for Simulation Models of Rural Residential Buildings Using Monthly Energy Bills: A Case Study in Zhejiang, China. *Case Stud. Therm. Eng.* **2025**, *73*, 106463. [[CrossRef](#)]
76. Zhang, Y.; Khan, A.A.; Zhao, W.; Xiao, X. Optimization of Cultivation Strategies Through Crop Yield Prediction for Rice and Maize Using a Hybrid CatBoost-NSGA-II Model. *Agriculture* **2026**, *16*, 423. [[CrossRef](#)]
77. Sarfarazi, S.; Mascolo, I.; Modano, M.; Guarracino, F. Application of Artificial Intelligence to Support Design and Analysis of Steel Structures. *Metals* **2025**, *15*, 408. [[CrossRef](#)]
78. Machine Learning | Google for Developers. Available online: <https://developers.google.com/machine-learning/decision-forests/overfitting-and-pruning?hl=pl> (accessed on 8 April 2026).
79. Hamad, K.; Alotaibi, E.; Zeiada, W.; Al-Khateeb, G.; Abu Dabous, S.; Omar, M.; Mantha, B.R.K.; Arab, M.G.; Merabtene, T. Explainable Artificial Intelligence Visions on Incident Duration Using EXtreme Gradient Boosting and SHapley Additive ExPlanations. *Multimodal Transp.* **2025**, *4*, 100209. [[CrossRef](#)]
80. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 1705.07874. [[CrossRef](#)]
81. Ma, T.; Zhang, C.; Ji, L.; Zuo, Z.; Beckline, M.; Hu, Y.; Li, X.; Xiao, X. Development of Forest Aboveground Biomass Estimation, Its Problems and Future Solutions: A Review. *Ecol. Indic.* **2024**, *159*, 111653. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.